# Fat and Fatter: Crash Risk and Retail Trading*

Qian Yang†

First Draft: January 6, 2021
This Version: October 7, 2021

**Abstract**

I estimate ex-ante crash probabilities and jackpot probabilities via novel machine learning methodologies. In particular, I introduce imbalanced learning techniques to facilitate rare events prediction. I then show that crash probabilities predict lower returns in portfolio and cross-sectional tests. One possible reason for the negative risk-return relationship is that at least a subset of retail investors (proxied by Robinhood traders) tend to buy high crash risk stocks, rendering them overpriced, and resulting in lower returns subsequently. Using Robinhood's introduction of commission-free option trading at the end of 2017 as a quasi-natural experiment, together with textual information from Reddit, I document causal evidence that retail participation significantly increases ex-ante stock crash risk. This effect is stronger for small firms.

*Keywords*: Crash Risk, Cross-Section of Stock Returns, Imbalanced Learning, Machine Learning, Robinhood, Tail Risk, Wallstreetbets.

*Previously circulated with the title "Fat and Fatter: Monthly Crash Risk and Investor Trading. I thank my advisors Naveen Khanna and Hao Jiang for their support and encouragement. I thank Kirt Butler, Nuri Ersahin, Ruslan Goyenko (discussant), William Grieser, Ryan Israelsen, Donghyun Kim (discussant), Tim Loughran, Sophia Li, Dmitriy Muravyev, Mark Schroder, Andrei Simonov, Parth Venkat, Dacheng Xiu, Hayong Yun, Morad Zekhnini, and participants in EFA Doctoral Tutorial 2021, 2021 International Risk Management Conference, 2021 Academy of Behavioral Finance & Economics, 2021 SoFiE Summer School with focus on machine learning, the 17th Annual Conference of the Asia-Pacific Association of Derivatives (APAD) and MSU brownbag for their valuable comments.

†PhD candidate, yangqia8@msu.edu, Eli Broad School of Business, Michigan State University.

# 1. Introduction

During the COVID-19 pandemic in early 2020, the stock market crashed and then quickly recovered in equally dramatic fashion. Pundits attribute much of this market volatility to speculation by retail investors, such as "Robinhood Traders" (Zweig, 2020). More recently, GameStop's stock price surged more than 1,700% following a coordinated short squeeze attempt by Reddit users. Market activity for this episode peaked on January 27th, 2021, when over 24 billion shares and 57 million options were traded on GameStop. These patterns have emerged more generally, as retail investors drive a larger fraction of equity and option volume, particularly during large market swings.[1] Considering this trend, I investigate whether retail investors contribute to tail risk in market behavior.

Existing literature has shown evidence that institutions are likely to be rational speculators, as they ride the bubble and earn abnormal profits (Conrad et al., 2014; Jang and Kang, 2019), when arbitrage might be costly (Pontiff, 1996; Shleifer and Vishny, 1997). An important assumption is that retail investors are more likely to be "noise traders" (De Long et al., 1990a), driven by sentiment and attention-grabbing events (Barber and Odean, 2000; Barber et al., 2020). When retail traders inflate the prices of lottery stocks, institutions are able to take advantage of retail trades, driving the bubbles even bigger, consequently precipitating crashes. However, literature is silent on any causal evidence as to whether retail trades can contribute to ex-ante tail risk. In this paper, I first estimate ex-ante monthly crash and jackpot probabilities via novel machine learning methodologies, and then use these probabilities as proxy for tail risk. I show that retail investors, proxied by "Robinhood Traders", tend to buy both high crash risk and high jackpot risk stocks, likely driving stock prices further away from their fundamental values, exacerbating potential price bubbles, and resulting in much lower returns in subsequent periods.

To show causal evidence that higher retail trading increases ex-ante crash risk, I exploit

[1]Estimates suggest that retail traders drive over 25% of total equity market volume (McCrank, 2021) and they are leveraging their positions with options more than ever (Banerji, 2021). Moreover, roughly 15% of retail traders are believed to be novice investors (Feuer, 2021).

a quasi-natural experiment. Robinhood introduced commission-free option trading at the end of 2017, providing a supply shock to inexpensive option trading opportunities, vastly increasing user adoption. The influx of retail traders likely made prices of both options and their underlying stocks more volatile, thus providing an ideal setting to study the causal effect. One problem in studying this event is that Robinhood did not start disclosing user data until 6 months after the event. To tackle this problem, I exploit a popular Reddit forum "Wallstreetbets", and scrape "regular" user comments that relate to option trading activities. This forum has been active since 2012. A further advantage of using Reddit posts for this experiment is that "regular" comments started to appear on this forum only at the end of 2017, coinciding with the event date, thus alleviating potential endogeneity issue. I classify stocks that Reddit users mention via both their ticker symbols and options related keywords as "treatment" stocks. Further, I provide evidence that Reddit comments are highly positively correlated with Robinhood users' trading behavior. This allows me to circumvent the data problem with Robinhood and identify stocks that are likely to have a high degree of retail trading from the time of introduction of commission-free option trading. I show that the ex-ante crash probability of such stocks increased by 0.11 standard deviations after the introduction, and the effect is more pronounced in small stocks, which could be due to limits to arbitrage (Diether et al., 2009; Chu et al., 2020).

Estimating ex-ante crash risk is important because though crashes are infrequent they can be extremely painful. Prior literature (Campbell et al., 2008; Conrad et al., 2014; Jang and Kang, 2019) employ logistic regressions to study extreme downside events at an annual frequency. Since we are interested in retail investors' trading behavior and their influence on crash risk, a shorter-term measure is necessary. Therefore I add to this literature by looking at monthly frequency, and define a "crash' if a stock's monthly log return drops by more than 20%, and "jackpot" if it increases by more than 20%. The reason for this cutoff is that the unconditional distribution resulting from this treatment is commensurate with prior literature: crashes and jackpots constitute approximately 5% respectively of all

observations.

One challenge of studying crashes is that the typical econometric toolkit is less ideal for forecasting rare events. Importantly, given that "crashes" and "jackpots" occur at very low frequency, this creates an "imbalanced sample" problem, where the "crashes" and "jackpots" have far smaller sample sizes as compared to "plain" cases. Using generic classification models such as logistic regression is likely to cause bias and substantively underestimate rare event probabilities (King and Zeng, 2001). One intuition for this argument is that the loss function in logistic regressions treats each individual observation equally, regardless of its class label, and thus the algorithm is largely minimizing the log loss from classifying the "plain" class, while the losses from "crashes" and "jackpots" are not given sufficient attention. Furthermore, the cost structure of misclassfication is asymmetric. Misclassifying a potential "crash" as "plain" case can cause substantial shareholder wealth loss, as compared to misclassifying a "plain" case as "crash".[2]

To alleviate this problem, I introduce the Synthetic Minority Over-Sampling Technique (SMOTE) (Chawla et al., 2002) designed for imbalanced classification problems. I apply SMOTE to create synthetic observations for "crashes" and "jackpots" via $K$-nearest neighbors, such that all three classes in the transformed data have similar sample sizes. When applying classification models on the transformed sample, the resulting loss function should pay sufficient emphasis on the log loss of "crashes" and "jackpots". Moreover, by using an $L-2$ penalized multinomial logistic regression (Ridge regression) coupled with validation, I tune the estimator to optimize out-of-sample (OOS) forecasting performance. Using OOS F1-scores as the classification criterion, I show that the performance for classifying both crashes and jackpots are substantially improved over the base logistic regressions. The OOS F1-scores increase many folds, from 0.05 to 0.13 for crashes, and from 0.02 to 0.11 for jackpots. Moreover, crashes and jackpots are more separated: the unconditional correlation between estimated crash and jackpot probabilities is approximately -25%.[3]

[2]See Chawla et al. (2002) for more detailed discussion for the asymmetric cost structure.

[3]Conrad et al. (2014) estimate a positive correlation between crashes and jackpots, and so do Jang and

Sorting stocks into decile portfolios based on estimates of monthly ex-ante crash probabilities, a zero-cost strategy long in high-decile crash risk portfolios and short in low-decile crash risk portfolios produces consistent and significant negative alphas, averaging approximately -1% monthly, at 1% level of statistical significance, across various factor models. The results are robust to both equal-weighting and value-weighting schemes. In subsequent Fama-MacBeth cross-sectional regressions, controlling for conventional and anomaly characteristics, including other tail risk measures such as $MAX$ (Bali et al., 2011), the coefficient on crash risk stays consistently and significantly negative throughout all specifications.

It is extremely difficult to distinguish between crashes and jackpots ex ante. Moreover, crashes and jackpots could occur in tandem, due to possible short-term reversals. Therefore it is reasonable to assume that retail investors, with limited resources and attention, could misclassify the two tails ex ante. Since retail investors are shown to have lottery preferences (Kumar, 2009), it seems logical to conjecture that they would mistakenly buy both high crash risk and jackpot risk stocks. If investors mistake high crash risk stocks for high jackpot risk stocks, then the high crash risk stocks would be overpriced, predicting lower returns in the subsequent periods, consistent with what is revealed in the data. More importantly, if retail investors overbought high crash risk stocks, exacerbating the individual stock price bubble, then the probability of its crash would be higher in the subsequent period, effectively rendering the already fat left tail even fatter.

To prove this conjecture, I combine two novel data sets to examine retail investors' trading behavior with respect to the two tails. The first is the Robintrack data that records the total number of users (shareholders) registered on Robinhood trading platform that are holding each individual stock, at approximately hourly frequency.[4] To match the frequency of estimated crash risk and jackpot risk, I construct a Robinhood trading measure as the

Kang (2019). Thus according to their estimates, stocks that have high probability of crash also tend to have high probability of jackpot in the same period. This leads to the question whether the models employed are able to classify the data successfully.

[4]Robintrack: `https://robintrack.net/`. On its website: "Robintrack keeps track of how many Robinhood users hold a particular stock over time. It generates charts showing the relationship between price and popularity, and compiles some lists using the data."

monthly change in log number of users for each stock. Then I regress this measure on crash risk, jackpot risk, and other firm characteristics. Consistent with prior literature, retail investors tend to buy stocks with high past returns, an attention-grabbing characteristic (Barber et al., 2020), as shown by the positive and significant coefficients on the $MAX$ measure. Importantly, I find that "Robinhood Traders" tend to buy both high crash risk and high jackpot risk stocks, even after controlling for other related characteristics including $MAX$, idiosyncratic volatility, and illiquidity. This suggests that retail investors' preference for lottery stocks leads them to mistakenly buy high crash risk stocks while chasing ultra-high returns. It is quite possible due to the extreme difficulty to distinguish between the two tails ex ante, given retail investors' lack of resources and attention.

Existing literature has shown causal evidence that retail trading contributes to increased stock volatility (Foucault et al., 2011). Whether retail investors can contribute to higher ex-ante crash probability of stocks is an open question. It has been suggested that retail investors are likely "noise traders", where they follow positive feedback trading strategy, induced by their sentiment over the past winners (De Long et al., 1990a,b). If retail investors chase past winners, then their action would push the stock price even higher in the short term. Consequently the subsequent probability of crash should be higher. To my best knowledge, this is the first study that provides causal evidence that higher retail participation does increase crash risk. At the end of 2017, Robinhood introduced commission-free option trading. In addition to its easy-to-use interface, the introduction provided a supply shock to retail participation. This event should provide an ideal experiment for examining retail influence on crash risk. Robinhood, however, does not disclose user trading data.

To circumvent the data limitation, I scraped comments posted on "Wallstreetbets", a "subreddit" and probably the most popular investment forum, to identify stocks with retail option trading activities after the shock, and classify them as treatment stocks. Through a difference-in-difference analysis, I show that on average treatment stocks with increased retail participation experience a significant increase in monthly ex-ante crash probability.

The crash probability of stocks with high retail investor participation increased by 0.11 standard deviations after the introduction. After the supply shock, investors are faced with higher deep out-of-money option prices, increased stock trading volume, and increased total volatility. The effects are stronger for smaller firms.

This paper makes several unique contributions to both crash risk and retail trading literature. To the best of my knowledge, this is the first study that jointly estimates ex-ante monthly crash and jackpot probabilities, and introduces imbalanced learning methodologies to improve forecasting performance for rare events. It is also the first to utilize a quasi-natural experiment to document causal evidence on the impact of retail trading on stock crash risk. Finally, it is the first study that combines Robintrack data and Reddit textual data to identify retail trading activities.

The paper is organized as follows. Section 2 discusses the contributions of the present study to existing literature. Section 3 provides summary statistics and describes key variables used in this study. Section 4 explains the prediction method and corresponding results for estimating monthly crash probabilities. Section 5 conducts asset pricing tests for crash risk in the cross-section of individual stocks; Section 6 analyzes retail investor behavior and their impact on crash risk. Section 7 displays various robustness tests. Section 8 concludes.

## 2. Contributions to Literature

The literature on crash risk is extensive in both corporate finance and asset pricing. On the corporate side, the literature is mostly concerned with the determinants of firm crash risk. These determinants are often motivated by managers hoarding bad news (Jin and Myers, 2006). The idea is that the hoarding delays the information transmission such that when it is ultimately released, there is a sudden drop of price corresponding to the size of the cumulative bad news. Motivated by this theory, the literature has proposed a list of determinants that could endogenously influence crash risk, such as earnings management

(Hutton et al., 2009), tax avoidance (Kim et al., 2011), annual report readability (Li, 2008), CSR (Kim et al., 2014), liquidity (Chang et al., 2016), short interest (Callen and Fang, 2015), and governance (Andreou et al., 2016; An and Zhang, 2013). Almost all of these determinants discussed in the literature can only be measured at an annual frequency, and thus are not suitable for the present study, which focuses on the short-term behavior of retail investors.

On the asset pricing side, there is a strand of option pricing literature that tries to extract information from option prices to determine the size of tail risk. Bates (1991) was among the early papers that study the relationship between option prices and crashes. They show that the 1987 stock market crash can be predicted by the unusually high prices of out-of-money S&P 500 futures put options. Further more, the paper indicates that the jump diffusion parameters implied by the option prices show that the crash could be expected. Pan (2002) provide theoretical support for the jump-risk premia implied by near-the-money short-dated options that help explain volatility smirk. Xing et al. (2010) study the relationship between implied volatility smirks and the cross-section of stock returns. They show that the difference between implied volatility of out-of-money put options and at-the-money call options show strong predicting power for future stock returns. Yan (2011) show that jump size proxied by the slop of volatility smile predicts cross-section of stock returns. More recently, Barro and Liao (2020) build a theoretical model that links the relative price of far out-of-money put options with the probability of rare macro disasters. They show that the relative price of far-out-of-money put options are positively associated with the probability of rare disasters, which they infer from monthly fixed effects in empirical test. This literature is primarily based on the theoretically motivated jump-diffusion model for asset prices, where return process is modeled as a linear combination of a diffusion process that is a standard geometric Brownian motion and a jump process. The size of the jump is frequently used to estimate tail risk probabilities. The present study differs from this literature, as it takes no stand on the underlying process of stock returns, but rather sets a fixed threshold for defining crashes

and jackpots, i.e. the left and right tails. The benefit of this procedure is that the definition is model free, and hence is invariant to model assumptions.

A third strand of literature on crash risk attempts to directly predict the probability of crashes. Chen et al. (2001) employs cross-sectional regressions to forecast skewness of daily stock returns. They show that negative skewness can be predicted by recent increase in trading volume and positive returns. Campbell et al. (2008) use a dynamic logit model to predict distress probabilities for the cross section of firms. They show that high-distress-risk stocks suffer from lower subsequent returns. Conrad et al. (2014) show that high-distress-risk stocks are also likely to become jackpots. They use a logit model to predict the probability of deaths and jackpots. Most recently, Jang and Kang (2019) exploits a multinomial logit model to jointly predict probabilities of crashes and jackpots at an annual horizon. They show that institutions appear to ride the bubble instead of trading against high crash risk stocks, and overpricing cannot be fully explained by investor sentiment. Since the estimation strategies in this literature are largely conducted at an annual frequency, the timing of crashes is highly uncertain, and thus not suitable for he present study, where our focus is on short-term behavior of retail investors. Moreover, these studies do not pass the test of new empirical asset pricing factor models. For example, Jang and Kang (2019) uses Fama-French three-factor model (Fama and French, 1993) plus a momentum factor (Carhart, 1997) as the benchmark. I show in Appendix B that, using more recent sample period from 1996 to 2019, while benchmarking against CAPM, Fama-French three-factor, and momentum augmented four-factor models, a replication of their zero-cost high-minus-low crash risk portfolios show significantly negative alpha. However, the alphas quickly turn economically and statistically insignificant when the five-factor model (Fama and French, 2015) is used.

Another important issue with this literature is a severe imbalanced sample problem. By definition, tail probabilities are very low compared to normal conditions, where the crashes and jackpots are extremely rare. As noted in Ripley (1996) and King and Zeng (2001), the poor finite sample properties in the imbalanced sample context would bias the coefficients, as

the majority class will be much better estimated than the minority class, while we care more about minority class than the majority class. The reason is two fold. First, it is the minority class that provides most incremental information. Second, the cost of misclassifying the minority class as the majority one is much higher than the opposite. In our case, if investors misclassify a crash-prone stock into a "plain" case, they would suffer huge monetary loss in the subsequent period for including this stock in their portfolios since the mean of portfolio return is lowered. On the contrary, if investors misclassify a "plain" case as crashes, it would be comparatively less costly. This asymmetric cost structure calls into question the loss function used in a typical classification method such as logistic regression. Since the generic loss function takes into account all observations, regardless of the class composition, in our case the "plain" cases would overwhelm the algorithm, leaving much less attention to the classification loss for "crashes" and "jackpots". The present study contributes to this literature by introducing novel machine learning technique "SMOTE" to address the issue of imbalanced sample problem, and shows that forecasting performance is greatly improved with this technique.

This study is also related to the literature on the relationship between investor trading and market efficiency and bubble formation. For example, De Long et al. (1990a), De Long et al. (1990b), and Abreu and Brunnermeier (2003) provide the theoretical and empirical evidence of positive feedback traders and their potential impact on market. Greenwood and Nagel (2009) show that inexperienced institutional investors might help the formation of bubbles. On the other hand, the literature shows that retail investors are by and large "noise traders" that could trade too much (Barber and Odean, 2000). Speculative retail traders tend to chase lottery-like stocks, experiencing subsequent negative trading alpha, and affect stock prices accordingly (Han and Kumar, 2013). Recent evidence from "Robinhood Traders" shows that they tend to herd more on extreme past-return stocks, which are more attention-grabbing (Barber et al., 2020), while there is also evidence that mimicking portfolios based on the characteristics of "Robinhood Traders" do not seem to underperform the market,

but instead could be a market stabilizing force (Welch, 2020). These seemingly conflicting results point to the difficulties in studying retail trading behavior due to data limitations. This present study is closely associated with this literature. First, I show consistent evidence that retail investors, proxied by "Robinhood Traders", do show preference for lottery-like stocks, manifested by their buying activity in both high-crash and high-jackpot probability stocks. Moreover, this paper contributes to the literature in showing that retail investors have no apparent ability to distinguish left and right tails ex ante, and thus their buying activity would potentially push the high crash risk stock prices high, predicting a lower return subsequently. More importantly, this paper utilizes a quasi-natural experiment to provide causal evidence that retail trading can contribute significantly to higher ex-ante stock crash risk. In this regard, Foucault et al. (2011) was one of the first papers that use a quasi-natural experiment to identify the causal effect of retail trading on stock volatility. This paper focuses on tail risk instead of the second moment.

Finally, this study is related to the emerging literature that studies the implications and applications of machine learning methodologies in asset pricing. The literature largely focuses on the "factor zoo" problem. For example, Kozak et al. (2020) applies a shrinkage method to construct an SDF that can summarize a large portion of the cross section of stock returns. Feng et al. (2020) introduces a novel machine learning methodology to test whether any new factor matters. Bianchi et al. (2021) applies machine learning techniques to estimate bond risk premiums. Another strand of this literature focuses on the forecasting power of machine learning techniques. For example, Gu et al. (2020) conducts a comprehensive study of various machine learning models and their superior power in estimating risk premiums, and shows large gains for investors. Erel et al. (2021) forecasts director performance via machine learning. The present study enriches this literature by introducing a novel technique from "imbalanced learning" methodologies to solve rare event prediction problems. The findings in this paper has many implications for future research in finance, especially in studying low probability events, such as firm bankruptcy and defaults and insurance policies. By applying

proper adjustment to the data, we can greatly alleviate the bias resulting from classifying rare events data, and thus improve the probability of avoiding misclassification of disaster events, increasing investor welfare.

# 3. Data

## *3.1. Variables*

For definition of crashes and jackpots, I use log monthly returns of -20% and 20% as the cutoff points. The choice is reasonable in the following sense. Prior literature uses log annual returns of -70% and 70% as the cutoff points. The unconditional probabilities of crashes and jackpots defined this way at the annual frequency are roughly 5%. My definition for monthly crashes and jackpots at the cutoff points of -20% and 20% agrees to this distribution. It follows that the dependent variables are defined as categorical, where $crash = -1$, $jackpot = 1$, and $plain = 0$. For independent variables, I use Compustat quarterly data to construct accounting variables, analogous to the annual measures used in Jang and Kang (2019), where I transform the frequency to short-term intervals to match the predicting task. These fundamental variables include: past three month market return, past three month stock excess return relative to the CRSP value-weighted market return, book-to-market ratio, asset growth, return on equity, total stock return volatility, total skewness, size, detrended turnover, firm age, tangibility, and sales growth. On top of these fundamental and stock return variables, I draw insight from option literature that shows predicting power of option pricing information for tail risks. I follow Xing et al. (2010) to construct the implied volatility smirk measure, which is defined as the difference between the implied volatility of out-of-money put option and the implied volatility of at-the-money call option. This measure is frequently used as a proxy for stock crash risk. Then I follow Barro and Liao (2020) to construct deep out-of-money relative option price measure, which

is motivated by their pricing equation for deep out-of-money put option:

$$\Omega = \frac{\alpha z_0^{\alpha} \cdot pT \cdot \epsilon^{1+\alpha-\gamma}}{(\alpha-\gamma)(1+\alpha-\gamma)} \tag{1}$$

Where $\Omega$ is the ratio of option price to implied stock forward price, and $p$ is the probability of a macro disaster event. Since the put option price implies extreme left tail event, then it follows naturally that the counterpart measure from call option price should contain information for extreme right tail events.

I use Option Metrics to construct these measures. Due to availability of option data, I limit my sample between the year 1996 and 2019. Following Xing et al. (2010) and Barro and Liao (2020), I perform the following screening for put options: 1) days to expiration between 10 and 180 days; 2) implied volatility between 0.03 and 2; 3) open interest greater than zero; 4) option price greater than \$0.125; 5) non-missing volume; 6) moneyness between 0.1 and 0.9. Analogously, to aid the joint prediction of jackpots, I also include the relative price of deep out-of-money call options, which obey the screening for put options, but with moneyness between 1.05 and 1.8. The option price is defined as the mean of offer and ask prices for each option contract. The relative price of a contract is the ratio between the option price and the implied forward stock price. I use open interest as weight to calculate a weighted-average relative price. Then I average the daily relative price for each month to construct a monthly measure for each stock. I require at least 10 days of available data for each month.

For return data, I use CRSP for daily and monthly stock returns and volumes. Following asset pricing convention, I require common stocks with share code of 10 or 11, and with stock prices greater than \$1 to avoid extreme outliers. For retail trading, I use Robintrack, which tracks Robinhood user holding of individual stocks. This dataset is available from May 2018 to August 2020. Finally, I scraped user comments from Reddit "Wallstreetbets" to conduct analysis on the impact of Robinhood introduction of commission-free option trading on retail

trading, and consequently its impact on stock crash risk. The comments are available from the end of 2017, coinciding with the introduction of commission-free trading, to the end of my sample. I will provide more detailed discussion of these two data sets in Section 6. Definitions of variables are in Appendix.

### *3.2. Summary Statistics*

The summary statistics for selected variables are shown in Table 1. I separately report the characteristics of "crashes", "plain" cases, and "jackpots" to examine their differences.

[Table 1 about here.]

As was discussed earlier, since our forecast horizon is one month, long-term explanatory variables might not be desirable as they may not account for regime change and hence lack sufficient flexibility (Elliott and Timmermann, 2016). Therefore, the variables are defined such that the longest lag used is one year in the case of sales growth, where I use quarter-on-quarter changes to account for seasonality.[5] All the other variables are lagged by less than 6 months, and in some cases, three months or one month.[6]

As shown in Table 1, we can discern some of the patterns to separate the three classes. For example, Crash and jackpot cases have smaller size and lower tangibility than plain cases. They tend to have lower past excess returns, and tend to happen when past three-month market return is low. They have lower detrended turnover, higher total volatility, higher total asset growth, higher sales growth, and higher skewness, while jackpots have the highest mean skewness among the three classes. Crash and jackpot stocks tend to be younger and value firms. Crash stocks tend to have negative mean return on equity, and jackpot stocks have high past ROE. Both crash and jackpot cases has higher SMIRK measure than the plain cases. Finally, deep out-of-money put and call option prices relative to underlying prices are

[5]For example, quarter one of this year on quarter one of last year.
[6]See Appendix for variable definitions.

far higher for crash and jackpot stocks, almost two times that of plain cases. These summary statistics are consistent with the findings in existing literature.

In the next section, I describe the estimation methodology for ex-ante crash and jackpot probabilities, and compare the baseline logit results and improved results of machine learning models.

# 4. Estimating Ex-Ante Monthly Crash Risk

In this section, I discuss the methodologies used in both the baseline model and improved machine learning models, and show comparisons of key performance metrics for out-of-sample forecasting.

## *4.1. Multinomial Logit Regression*

As a precursor to out-of-sample predictions, I first run an in-sample multinomial logit regression to examine whether the selected variables are strongly correlated with future realized crashes and jackpots, and whether the model is economically sound. Table 2 shows the results. Standard errors are clustered at both firm and time levels per Petersen (2009).

[Table 2 about here.]

All independent variables are standardized, so that we can directly compare the importance of each variable with others. Table 2 shows that the relative option prices of deep out-of-money puts and calls are significant predictors of crashes and jackpots in next month. Though they have the same positive sign, the coefficient on put options for crashes are greater than that for jackpots, while the coefficient on call options for crashes are less than that for jackpots. This makes intuitive sense: high relative price for deep out-of-money put options signals greater demand for protection for that particular stock, which precedes the pending crash; high relative price for deep out-of-money call options signals greater demand

for speculation for that particular stock, which precedes the pending jackpot. On the other hand, surprisingly, the implied volatility SMIRK measure shows no significance in predicting crashes and jackpots. All the other variables show coefficients in signs that are largely consistent with literature. Importantly, the size of the coefficients show that only size, volatility and age have comparable importance to option variables. This exercise shows that the model has substantial explanatory power with these selected variables. Next, I move on to discuss the machine learning models used to predict ex-ante crash and jackpot probabilities.

## *4.2. Out-of-Sample Forecasting via Machine Learning*

### *4.2.1. $L-2$ Penalized Multinomial Logistic Regressions*

The in-sample logit shows significant explanatory power. However, it is well known that in-sample fit has substantial overfitting problem that leads to poor out-of-sample performance.

To address this issue, I follow prior literature and conduct a rolling window estimation procedure, where I use 6 months of data as the training sample and 1 month data as the test sample in each window. For example, the first window consists of training sample from January 1996 to June 1996, and test sample of July 1996; the second window consists of training sample from February 1996 to July 1996, and test sample of August 1996, and so on. This procedure produces true out-of-sample estimates of crash and jackpot probabilities for next month.

To improve the forecasting power and address the overfitting issue, I apply logistic Ridge regression as my main model.[7] There are several reasons that Ridge is chosen. First, it is an extension of logistic regression, and hence relatively easy to interpret. The model produces interpretable coefficients. We are able to tune the model by the penalty factor $\lambda$ to search

[7]In Section 7 and Appendix, I explore two set of tests: one set uses the same underlying variables as the main test, but use other machine learning models; the other set expands the variables to 134, transforms the data via PCA, and then applies ridge regression, XGBoost and neural networks. I show that using either alternative algorithms or more complex models do not materially change the results. Hence for simplicity and interpretability, I present the simple ridge regression as the main model.

for the best estimator with respect to out-of-sample performance. The multinomial logistic Ridge seeks to estimate:

$$Pr(G = k|X = x) = \frac{\exp \beta_{0k} + \beta_k^T x}{\sum_{l=1}^{K} \exp \beta_{0l} + \beta_l^T x} \tag{2}$$

Where K is number of classes. The general elastic net (Zou and Hastie, 2005) penalized negative log-likelihood function can be written as:

$$\begin{aligned} \ell(\{\beta_{0k}, \beta_k\}_1^K) = -[\frac{1}{N} \sum_{i=1}^{N} (\sum_{k=1}^{K} y_{il}(\beta_{0k} + x_i^T \beta_k) - \log (\sum_{l=1}^{K} \exp \beta_{0l} + x_i^T \beta_l))] \\ + \lambda[\frac{1}{2}(1-\alpha)\|\beta\|_F^2 + \alpha \sum_{j=1}^{p} \|\beta_j\|_q] \end{aligned} \tag{3}$$

Where $\lambda$ is the penalty factor for the weighted L-1 and L-2 penalties, $\alpha$ is the weight of L-1 penalty. Hence Ridge regression is a special case when $\alpha = 0$. $L-2$ penalty is particularly suitable in our setting, since model sparsity is not a concern (number of variables are far less than number of observations).

In each rolling window, I split the training sample into two parts: first 5 months as the training set, and the last 1 month as the validation part.[8] Then I use the training set to tune the penalty factor $\lambda$ of the Ridge model, and use the validation set to find the best Ridge estimator. Then this estimator is used to fit the test sample in the same window. This rolling window data split scheme can be represented in Figure 1.

[Fig. 1 about here.]

As shown in the figure, the bars represent months of data in a window. From top to bottom: the first bar represents the training set, which consists of five months of data; the

[8]Cross validation is usually used in training machine learning models, where the data is assumed to be i.i.d. However, in this setting, the data is in a panel structure, where observations might show substantial temporal structure. This time structure contains valuable information, and hence generic cross validation would ignore this structure and hence produce possibly inferior results. See for example Roberts et al. (2017) for detailed discussion. Nonetheless, in robustness tests, I show that by using three fold cross validation, the results are not materially altered.

second bar is the validation set, consisting one month of data; the last bar is the test set, consisting one month of data. For example, the first rolling window consists training data (January 1996 to May 1996), validation data (June 1996), and test data (July 1996). The training set is used to fit the model; the validation set is used to tune the hyper parameters to find the best estimator in terms of forecasting metric (e.g., F1 score); the resulting optimal estimator is then used to fit the test data, and compare the prediction with the ground truth, in order to generate the test performance metrics. As a comparison, I also apply the simple logit model on the whole training sample (6 months), and use the coefficients to directly fit the test sample.

#### *4.2.2. Imbalanced Learning*

Since crashes and jackpots are rare events with unconditional probabilities of occurrence at less than 5%, the usually logistic estimator would produce biased estimates due to the poor finite sample properties.[9] I provide a simple intuition for this argument. The cost of misclassifying either crashes or jackpots as "plain" cases is far higher than misclassifying "plain" cases as crashes or jackpots. If the former situation happens, investors are either faced with huge unexpected losses or missed opportunities, whereas the latter would be analogous to giving up average returns. Thus, the cost of misclassification is asymmetric. On the other hand, the loss function in a generic logit regression is not cost sensitive, meaning that it treats each observation equally.

For simplicity, let's only consider two classes: "crashes" and "plain" cases. When using logistic regression, its loss function is log loss, or cross-entropy, as represented by Equation 4.

$$logLoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{4}$$

Now we separate the two classes and denote the sizes of them as $N_{plain}$ and $N_{crash}$, where $N_{plain}$ denotes the number of "normal" observations, and $N_{crash}$ denotes the number

[9] See for example King and Zeng (2001) for discussion.

of crashes. Then the log loss function can be written as:

$$logLoss = -\frac{1}{N_{plain} + N_{crash}}[\sum_{i=1}^{N_{plain}} \log(p_i^{plain})] - \frac{1}{N_{plain} + N_{crash}}[\sum_{i=1}^{N_{crash}} \log(p_i^{crash})] \quad (5)$$

Where the first term refers to the log loss of classifying "plain" cases, and the second term refers to that of "crashes".

Now consider the "imbalanced sample" case, where $N_{plain} >> N_{crash}$. In the extreme case, consider fixed $N_{crash}$ and $N_{plain}/N_{crash} \to \infty$. Then the second term of Equation 5 tends to zero, and effectively we are only minimizing the log loss on the "plain" cases. King and Zeng (2001) shows that in finite sample, using generic logistic regression on imbalanced sample, or "rare event classification" problems, would produce biased coefficients and underestimate the probability of rare events. The argument can be easily extended to cases of multiple classes.

To address this issue, I introduce a widely used machine learning technique that is novel to finance literature: Synthetic Minority Over-sampling Technique (SMOTE), introduced in the seminal paper by Chawla et al. (2002). Oversampling is achieved by creating synthetic observations along the lines in the feature space that join the minority class $K$-nearest neighbors.[10] More formally, let $X_{minority}$ be an observation of minority class observed in the training sample, let $\widetilde{X}_{minority}$ be a random neighbor sampled adjacent to $X_{minority}$. Then a synthetic minority observation can be generated as in Equation 6:

$$X_{minority}^{syn} = w \cdot X_{minority} + (1 - w) \cdot \widetilde{X}_{minority} \quad (6)$$

Where $w \in (0, 1)$ is a random number. The $k-$nearest neighbors are sampled repeatedly with replacement, and corresponding synthetic observations are created until the desired balance between minority and majority classes are achieved.[11]

[10] See Friedman et al. (2001) for introduction to $K$-nearest neighbors.

[11] In our case, balance means that crashes, jackpots, and plain cases have the same number of (synthetic and real) observations.

The intuition is that assuming the features (characteristics) of the minority class are sufficiently clustered, it is reasonable to create "similar" observations within that cluster. In this paper, in order not to lose information of the majority class, I use SMOTE to create synthetic observations for both crashes and jackpots using oversampling, while keeping all "plain" examples without under-sampling. Since all classes are balanced, the loss function would pay equal attention to the log losses of all three classes, thus alleviating the "imbalanced sample" problem. I show that using this technique can greatly improve the metrics for crashes and jackpots.

#### *4.2.3. Metrics and Results*

I follow machine learning literature and choose the following metrics: precision, recall, and F1-score (Seliya et al., 2009). The common accuracy measure that is used in most forecasting literature is not suitable in imbalanced sample classifications (Batista et al., 2004). Therefore I do not report accuracy. The definitions for precision, recall, and F1-score are as follows:

$$Precision = \frac{True\,Positives}{True\,Positives + False\,Positives} \tag{7}$$

$$Recall = \frac{True\,Positives}{True\,Positives + False\,Negatives} \tag{8}$$

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

These metrics are computed for each of the three classes: crash, plain, and jackpot. Since I use rolling window estimations, each rolling window exercise can generate a set of metrics that evaluate out-of-sample performance, then I compute the mean metrics. There are in total 281 rolling windows and the same amount of associated sets of metrics. I summarize the mean metrics for simple logit and Ridge in Table 3.

[Table 3 about here.]

Table 3 show that across the board, especially in crash and jackpot categories, Ridge

regression shows far superior performance than the simple logit. For example, the recall of crashes on average improves by a factor of nearly 70, while the F1 score of crashes on average improves by a factor of around 2.5. In the case of jackpots, the recall improves by a factor of 25, while the F1 score improves by a factor of 6. It is important to note that precision for both tail classes suffer a bit, though if looked at alone, they are misleading in that it only cares about how many observations are true out of all the predicted observations in that class. In our case, the recall measure is more important, as it identifies the model's ability to capture the true classes as much as possible. F1 score seeks to balance the two measures, and provides a more nuanced view of the model's power.

To visually demonstrate the comparison of metrics between models, I also plot the confusion matrices for the two models in the aggregate sense, where I simply add up predicted classes across time. A confusion matrix is a square matrix, where the rows are designated as true classes, and the columns are designated as predicted classes. Hence the diagonal elements are true classes that are successfully predicted. Then it follows that if we normalize the matrix row by row, the diagonal elements can be viewed as recall for each class. Figure 2 plots the matrices for all models.

[Fig. 2 about here.]

As shown in Figure 2, the Ridge model performs substantially better than the simple logit, as it is shown that The gravity of each class is more heavily concentrated along the diagonal, which is more ideal.

As an illustration of the predicted probabilities, I plot the monthly mean crash and jackpot probabilities in Figure 3.

[Fig. 3 about here.]

On top of the improved out-of-sample performance, the Ridge model seems to separate the left and right tails pretty well: the unconditional correlation between crash and jackpot

probabilities is around -25%, while prior studies (for example, Conrad et al. (2014)) often show strong positive correlation between the two tails. Overall, machine learning models combined with SMOTE produce far superior out-of-sample results as compared to simple logit.[12]

To further understand what variables have relatively large effect on predicting crash and jackpot probabilities, I plot coefficients for crash and jackpot classes. Since we have 281 windows, we have 281 sets of coefficients for both classes. We want to understand the relative importance of each feature, hence I obtain the absolute value of each coefficient. All variables are standardized before training, and thus the absolute values represent relative "size" of their importance. I plot them in two separate heat maps as in Figure 4 and Figure 5.

[Fig. 4 about here.]

[Fig. 5 about here.]

As each heat map plots the size of the coefficients with different degrees of depth of colors, the deeper the color, the larger the absolute value of the coefficient, and hence the importance. The two figures have very similar patterns across time, meaning that the variables have consistently relative importance across classes and time. Importantly, the two option variables stay relatively important throughout the periods, thus contributing significantly towards the forecasting power of the model. Another interesting result to note is that the variable $RM3$ seems to have the deepest colors across classes and time. This is not surprising, as $RM3$ is the excess return of the market in past three months. Our target responses, "crash" and "jackpot", are defined by setting a fixed set of thresholds, thus do not distinguish between systematic and unsystematic part of the returns. When market experiences extreme returns, individual stocks are also likely experiencing extreme returns due to their exposure to market risk. Therefore in this estimation procedure, market excess return is a

[12]I show in Section 7 and in Appendix that more complex models can be used; however, they produce similar results. Thus the simple model is presented as main result for its simplicity and interpretability.

strong predictor. An interesting question would be that what if we separate systematic and unsystematic part of return and try to see if we can successfully predict the idiosyncratic extreme returns? This is beyond the scope of this paper, but one thing might be certain that it would be extremely hard to predict idiosyncratic tail risks, since by definition and assumption of the underlying asset pricing models, idiosyncratic tail risks are unpredictable. Of course, in reality, we are never sure if we can successfully tease out all systematic part of the returns. As shown in Herskovic et al. (2016), idiosyncratic volatility has strong factor structure, likewise, the idiosyncratic tail risks likely also have factor structure that one can exploit. This is an interesting path for future research that is beyond the scope of this paper.

Next, armed with a set of estimates that can more reliably predict crashes and jackpots, I turn to implications of monthly crash risk for cross-section of stock returns.

# 5. Are Monthly Crash Risk Priced?

In this section, I examine whether ex-ante monthly crash risk is priced in the market. Literature has long shown that tail risk is priced, as investors have hedging demand against extreme tail events.[13] Prior studies such as Conrad et al. (2014) and Jang and Kang (2019) show that the two tails are likely negatively priced, as investors follow positive feedback strategies, which renders these lottery like stocks overpriced, and they subsequently experience lower returns. While they focus on the next year's crash risk, this paper studies more short-term phenomenon, where I estimated firms' ex-ante monthly crash risk, jointly with jackpot risk. Following the same rationale, we would expect these risks to be priced in the market. I proceed first in a portfolio test, and then examine the issue in the cross section.

[13]See for example Kelly and Jiang (2014).

### *5.1. Time-Series Portfolio Tests of Monthly Crash Risk*

I run time-series portfolio return regressions on time-series factors benchmarking various asset pricing models. The asset pricing models include: CAPM market model, Fama-French three-factor model (FF3) (Fama and French, 1993), then augmented with a momentum factor (FF4) (Carhart, 1997), Fama-French five-factor model (FF5) (Fama and French, 2015), and finally FF5 augmented with momentum factor (FF6). At the end of each month, I sort the stocks based on their predicted next-month crash probabilities from the Ridge model into decile portfolios, then I calculate either equal-weighted or value-weighted portfolio returns for the top decile and bottom decile, and form a zero-cost trading strategy by longing the top decile and shorting the bottom decile, and regress the excess returns on pricing factors. I apply common asset pricing filters to the stocks: stocks with a share code of 10 or 11, and month-end price of greater than \$5. The results are shown in Table 4.

[Table 4 about here.]

As shown in Table 4, when we long highest crash risk decile portfolio and short lowest decile portfolio, we produce consistent and significant negative alphas across different asset pricing models, equal-weighted or value-weighted, with $t$-statistics of magnitude of well over 3. Next I show more detailed results for the ten value-weighted decile portfolios, to examine the return behavior of crash risk. The results are shown in Table 5.

[Table 5 about here.]

As shown in Table 5, Panel A reports the resulting alpha estimates from regressing excess portfolio returns on various asset pricing factors, from simple one-factor CAPM model all the way up to most recent Fama-French five factor model. The value-weighted decile portfolio alphas largely decrease monotonically from bottom decile in crash risk to top decile. It shows that monthly crash risk is negatively priced, consistent with the results from using annual measures of crash risk in prior literature. Panel B reports the coefficients from regressing

the ten decile portfolio excess returns on Fama-French five factors plus a momentum factor. Again the alphas largely decrease from the bottom to top decile portfolios. The beta estimates on the factors make intuitive sense: higher ex-ante crash risk portfolios tend to have higher market beta, suggesting they are more risky and have higher exposure to market risk; high crash risk portfolios tend to concentrate in small firms, as evidenced by the large and positive loadings on SMB factor; they tend to be growth firms, with lower profitability, and low past year returns; the loadings on CMA is not very significant, however the sign suggests that high crash risk firms tend to have less aggressive asset growth. These results conform with the multinomial regression results we obtained when estimating ex-ante crash risk. Taken together, there is strong evidence that this monthly ex-ante crash risk is priced in the market, and is negative correlated with future returns.

### *5.2. Monthly Crash Risk and Cross-Section of Stock Returns*

Next, I examine the relationship between monthly crash risk and cross-section of stock returns. I run Fama-MacBeth regressions (Fama and MacBeth, 1973) following the procedure in Fama and French (2020), where I regress raw stock returns on cross-sectionally standardized lagged firm characteristics. Then the coefficients on characteristics can be directly interpreted as average priced return spread of one standard deviation of the corresponding firm risk. I include common risk characteristics such as size, book-to-market ($B2M$), asset growth ($ATG$), profitability ($ROE$), momentum ($MOM$), short-term reversal ($REV$), and my estimated crash probabilities and jackpot probabilities. On top of these variables, I follow Jang and Kang (2019) and control for a battery of anomaly characteristics that are shown to be significantly correlated with future stock returns: abnormal capital investment $ACI$ (Titman et al., 2004), illiquidity $ILLIQ$ (Amihud, 2002), turnover $TURN$, idiosyncratic volatility $IVOL$, asset growth $AG$ (Cooper et al., 2008), composite equity issues $CEI$ (Daniel and Titman, 2006), gross profitability $GP$ (Novy-Marx, 2013), net operating assets $NOA$ (Hirshleifer et al., 2004), net stock issues $NSI$ (Ritter, 1991), and O-score $OSCR$

(Ohlson, 1980).

Importantly, Bali et al. (2011) propose a measure $MAX$ that represents investors' preference for lottery like stocks. $MAX$ stands for the maximum daily return achieved by each stock in the prior month. To see if my measure carries additional information that distinguishes from $MAX$, I add $MAX$ measure as a control variable in the Fama-MacBeth regressions. I report the regression results in Table 6.

[Table 6 about here.]

Table 6 show that even after controlling for common risk characteristics and a plethora of anomaly variables, and finally the $MAX$ measure, the loadings on ex-ante monthly crash risk remains economically and statistically significant, with comparable magnitude with the time-series portfolio alpha results. One-standard-deviation change in ex-ante monthly crash risk predicts negative return spread between -0.308% to -0.251%. The coefficient on jackpot risk is positive and significant at conventional statistical levels, consistent with the findings in Jang and Kang (2019). These results provide strong support for the efficacy of the prediction model combining multinomial ridge regression and SMOTE, and show consistent evidence that ex-ante monthly crash risk is robustly priced in the market.

### *5.3. Where Are the Arbitrageurs?*

Given the large and negative return spread between the bottom decile and top decile ex-ante crash risk portfolios, an obvious question is why this is not arbitraged away. After all, if market is efficient, such opportunities would be quickly taken advantage of by rational arbitrageurs, and hence on average we should not be able to see consistent alphas. One possibility is limits to arbitrage (Shleifer and Vishny, 1997), that it might be prohibitively costly to short the high ex-ante crash risk stocks. To shed some light on this issue, I examine the mean characteristics of each decile portfolios. I'm especially interested in size, MAX, idiosyncratic volatility, and illiquidity. Table 7 reports these simple statistics.

[Table 7 about here.]

Table 7 gives a simple and clear picture that, from bottom decile to top decile crash risk portfolios, size goes monotonically down, while $MAX$, idiosyncratic volatility, and illiquidity go monotonically up. This is not surprising, since the coefficients on these variables are consistent with the multinomial logit regression when we try to estimate the crash risk. Nevertheless, this shows that the higher the ex-ante crash risk, the more difficult it might become to short the stock. The following simple test would reinforce this notion. I run a panel regression of monthly short interest on ex-ante crash risk, jackpot risk, and other lagged stock characteristics. Short interest is defined as adjusted shares sold short scaled by total shares outstanding for each stock. All variables are standardized cross-sectionally to be mean zero and standard deviation of one so that we may interpret the results easily. To control for possible unobserved heterogeneity, I include both firm and time fixed effects. The results are reported in Table 8.

[Table 8 about here.]

Table 8 shows some interesting results. First, from Column (1), there seems no concrete evidence that crash risk is being actively shorted, when controlling for other firm characteristics. However, in Column (2), when we add an interaction term between crash risk and size, the loading on crash risk turns positive and significant at statistical level of 1%. Note that since all variables are standardized, this means that a one standard deviation increase in ex-ante crash risk is associated with 0.02 standard deviation increase in short interest ratio. More interestingly, when we look at the interaction term, if the stock is one standard deviation below the mean size, while one standard deviation above mean crash risk, the netted effect is -0.025 standard deviation decrease in short interest ratio. On the flip side, if the firm is a large firm with high crash risk, it is more likely to be shorted.

These results seem to convey a simple message, that since high ex-ante crash risk stocks tend to be small with low liquidity, they are difficult to short, and in reality indeed they are

less likely to be shorted even when there is money left on the table. Hence this speaks to the result that the negative alpha on the top decile portfolio remains economically large and consistent. To further explore the issue of how investors perceive crash risk and trade this risk, next I turn to institutions and retail investors to explore their trading behavior with respect to the left tail.

# 6. The Impact of Retail Trading on Crash Risk

Prior literature primarily focuses on the institutional trading behavior with respect to stock crash risk. For example, Conrad et al. (2014) and Jang and Kang (2019) show evidence that institutional investors tend to "ride the bubble" as rational speculators, instead of trading against crash risk as rational arbitragers. They argue that such behaviors may drive the stock prices further away from fundamentals, exacerbating the bubble conditions à la De Long et al. (1990a), De Long et al. (1990b), and Abreu and Brunnermeier (2003).

On the other hand, retail investors are assumed to be "noise traders" that could trade too much (Barber and Odean, 2000), and those speculative retail traders tend to chase lottery-like stocks, experiencing subsequent negative trading alpha, and affect stock prices accordingly (Han and Kumar, 2013). Recent evidence from "Robinhood Traders" show that they tend to herd more on extreme past-return stocks, which are more attention-grabbing (Barber et al., 2020), while there is also evidence that mimicking portfolios based on the characteristics of "Robinhood Traders" do not seem to underperform the market, but instead could be a market stabilizing force (Welch, 2020). In summary, literature shows that as price takers, retail investors chase lottery-like stocks, which is consistent with theory.

However, none of these studies touch on the impact of retail trading on crash risk. First, prior literature usually groups left tail and right tail together, and studies the relationship between investor trading and lottery-like characteristics of stocks. Therefore, there is no clear research on how retail investors trade with respect to the two tails. Given the difficulty

of separating the two tails ex ante, it is reasonable to assume that retail investors, with limited resources and attention, would find it hard to distinguish high crash risk stocks and high jackpot risk stocks. Following this confusion, if retail investors have a preference for lottery characteristics, they would tend to buy both tails for lack of perfect foresight. Second, recent episodes of retail trading introduced in the first page of this present study show that retail investors can be marginal price setters under certain circumstances. However, there has been scant research that shows causal evidence for this conjecture.

In this section, I first look at trading behavior of retail investors with respected to the estimated ex-ante crash and jackpot probabilities. Then I explore a quasi-natural experiment to infer the causal effect that retail investors do tend to increase ex-ante stock crash risk.

### *6.1. Retail Trading of Crash Risk*

To examine the trading behavior of retail investors, I construct retail trading imbalance measure from Robintrack data. As has been extensively discussed in Barber et al. (2020) and Welch (2020), Robintrack data contains hourly stock popularity numbers that are measure by how many users on Robinhood hold a particular stock at certain hour. Since we cannot observe the number of shares they hold for each stock, and there is no data for total number of users for each time period, the next best thing we can do is to measure the change in number of users for each stock. As my risk measures for crashes and jackpots are estimated at monthly frequency, I use month-end numbers of Robinhood users to merge the data. Therefore the measure for retail trading can be constructed as in Equation 10:

$$Change\#User = \log(\#User_{i,t}) - \log(\#User_{i,t-1}) \tag{10}$$

Where $t$ is at monthly frequency.

With this measure, I now explore their trading behaviors. The Robinhood sample runs from May 2018 to November 2019, subject to data limitation. On top of ex-ante monthly

crash risk and jackpot risk measures, I control for common risk characteristics, which include size, excess return over the market over the last quarter, detrended turnover over the last quarter, asset growth rate over the last quarter, tangible assets, sales growth, ROE of the most recent quarter, firm age, and book-to-market ratio. I also add the following variables as additional controls: betas of Fama-French 3-factor models by running daily regressions of excess returns on factor returns over the last quarter, idiosyncratic volatility as the residual volatility obtained from the above regressions, and total volatility of the stock over the last quarter. Finally I add $MAX$ (Bali et al., 2011) measure as a control for lottery characteristics.

I now examine the trading behavior of retail investors proxied by Robinhood traders, using the imbalance measure inferred from Robintrack data. Following the prior procedure, but at a monthly frequency, I first run Fama-MacBeth cross-sectional regressions of retail trading imbalance on ex-ante monthly crash and jackpot risks, controlling for other characteristics, and then run a panel regression, where I add firm and time fixed effects to control for unobserved heterogeneities. The results are shown in Table 9.

[Table 9 about here.]

If retail investors are able to perfectly distinguish between left and right tails ex ante, we should see a negative coefficient on crash risk and a positive coefficient on jackpot risk. However, Table 9 shows consistently that the coefficients on both crash and jackpot risks are positive and significant in both Fama-MacBeth regressions and panel regression, suggesting that Robinhood traders are buying both high crash risk stocks and high jackpot risk stocks, consistent with prior literature that they have a preference for lottery-like stocks. Importantly, the regressions control for $MAX$, another proxy for lottery characteristics, which shows that the ex-ante crash risk and jackpot risk capture additional information about retail investor preferences. Moreover, the significant and positive loadings on $MAX$ provide evidence that is consistent with literature that retail traders tend to buy attention-grabbing stocks (Barber et al., 2020). In summary, their buying activities would likely push both

high crash risk and high jackpot risk stock prices high, and subsequently leading to negative returns in the next month, as shown in previous pricing tests.

### *6.2. The Impact of Retail Trading on Crash Risk*

We have established evidence that retail investors, as proxied by Robinhood traders, seem to display a strong preference for high ex-ante crash and jackpot risk stocks, even after controlling for other characteristics and $MAX$. Since retail traders are generally considered as "noise traders", their trading activities would logically add noise to stock return distribution. Then it seems logical to reach a conjecture that more retail trading would lead to increased ex-ante crash risk. This section explores this question.

There has been much debate in literature whether and how much retail investors can affect stock prices. Classical asset pricing models assume rational investors are price takers, and there is no room for price impact (Merton, 1973). Recent evidence suggests that retail investors do affect stock volatility (Foucault et al., 2011). They may be marginal price setters for small stocks (Graham and Kumar, 2006). Retail short sellers predict negative future returns, and they seem to have superior knowledge of small firm fundamentals (Kelley and Tetlock, 2017). Much of the literature focus on predictive tests, as it is extremely difficult to find ideal settings for proper identification for any claims for causality. Foucault et al. (2011) was one of the papers that use quasi-natural experiments to identify the causal effect of retail trading on stock volatility.

Another strand of literature that is relevant to this study is the feedback effect between option trading and stock trading, as two significant predictors of crash and jackpot risks are far-out-of-money put and call option relative prices with respect to stock forward price. Anthony (1988) was among the first to examine the sequential information flow from options to stocks. Hence the deep out-of-money options themselves are good proxies for ex-ante stock crash risk. Therefore, in subsequent tests, I look at both predicted crash risk and the deep out-of-money option variables.

I explore a quasi-natural experiment: Robinhood introduced commission-free option trading on its platform on December 12, 2017, which would take effect in 2018 (Robinhood, 2017). Even today, option trading is generally not free on other platforms.[14] While the option trading fees are declining in recent years partly perhaps due to Robinhood, the fees today still ranges around $0.65 per contract.

After the introduction of commission free option trading, Robinhood traders appear to have developed a zeal for option trading, so much that they actively discuss their Robinhood positions and gains and losses on social platforms, especially on Reddit. After all, option trading brings the benefit of cheap leverage that enables them to bet big with relatively small amount of money. In fact, around 13% of Robinhood users trade options, according to the firm disclosure.[15] This is not a small number, considering the total users amount to 13 million in 2020, and hence there are at least 1.69 million users on Robinhood actively trading options.[16] This influx of Robinhood option trader army should drive the demand for options for popular stocks, and thus affect option prices. The trading of popular stock options should in turn transmit to the elevated trading activities in the underlying stocks. This event was not caused by underlying option or stock returns or volatilities, and hence should serve as a suitable experiment.

Based on prior analysis, I hypothesize that after the introduction of commission-free option trading, those stocks whose options experienced influx of Robinhood traders should observe their ex-ante crash risk increasing, compared to similar stocks that do not have this influx around the event. One source of the increase might come from increased demand of deep out-of-money options. This increased trading of options should in turn translate into increased trading of the underlying stocks. One difficult issue, however, is that there is no direct way to identify which stocks experienced influx of Robinhood traders with respect to their options. Even though we do observe which stocks are popular among Robinhood

[14]See for example an article comparing option trading fees for all major discount brokerage houses on `https://www.bankrate.com/investing/best-brokers-options-trading/`.

[15]See article McCabe (2020).

[16]See `https://www.businessofapps.com/data/robinhood-statistics/`.

traders, but unfortunately Robinhood do not share their option trading data.

To circumvent this issue, I explore textual information from the popular online social media platform "Reddit" and its particularly popular subreddit "WallstreetBets".[17] As of January 2021, this subreddit has 1.8 million total active users, who post regularly everyday. I explore two "flairs" in this subreddit: "daily discussions" and "what's your move tomorrow?". I choose these two flairs because users post here every trading day, such that I have a steady number of posts and comments. I scraped all the first-level and second-level comments each day from December 2017 to September 2020. These comments are short in nature, with colorful languages. I perform two layers of pre-processing: first, I find out all the posts that contain valid ticker names. I discard those tickers that are also common English words, slangs, or month abbreviations (e.g. SEP). Second, I find out all the posts with tickers that mention "option", "call', or "put" to identify possible option buying activities. I assume that, if a user posts a comment with tickers in it, and mentions option terms in the same post, then he/she is more likely to have traded in these options, which is a reasonable assumption. Through this methodology, I can identify which stocks are likely to experience sudden influx of retail traders with respect to both options and underlying stocks.

To illustrate the extent to which they mention stocks and options in their comments, for each day in the sample, I summarize the number of unique posts that contain tickers, of which number of posts that mention options, number of unique firms mentioned, of which number of firms that mention options. I then plot the two series as in Figure 6 and Figure 7.

[Fig. 6 about here.]

[Fig. 7 about here.]

Subsequently, I use the firms that are co-mentioned with options in Wallstreetbets comments as a proxy that retail investors participate in the option trading of these stocks after

[17] *Wallstreetbets*: `https://www.reddit.com/r/wallstreetbets/`.

Robinhood introduction of commission-free option trading in December 2017. Therefore, the sample can be divided as following: I restrict my attention to the year 2017 and 2018, with 2018 as post event period. The aforementioned firms will be the treatment group, and the rest with valid crash probability estimates as the control group.

Before the actual estimation begins, we must examine whether Reddit mentioning is a reasonable proxy for retail trading activities. If this assumption is true, then the stocks whose ticker and option keywords are mentioned in Wallstreetbets comments would experience higher trading and user popularity on Robinhood. To examine this assumption, I focus on the sample period where I observe both Reddit comments and Robintrack data, which is between May 2018 to December 2019. Note that our experiment happened at the end of 2017, and thus this test is outside of that event period.

Then I regress the following dependent variables on a dummy "WSB Option Flag" and a set of control variables: Robinhood trade imbalance, measured by the change in log of number of users; log of number of users, as a proxy for the popularity of a certain stock; trading volume; and change in trading volume. All variables are measured at daily frequency to take advantage of the data. I use both Fama-MacBeth regressions and panel regressions to examine the loadings on the variable of interest: "WSB Option Flag", which means on day $t$, stock $i$ is mentioned together with either "option", "call", or "put" in the same comment. I report in Table 10 that in all specifications, the loadings on "WSB Option Flag" is positive and highly statistically significant. This provides evidence that Wallstreetbets mentioning is an effective identifier of retail participation.

[Table 10 about here.]

Following the aforementioned reasoning, I conduct a standard difference-in-difference analysis similar to that in Foucault et al. (2011). I estimate the following equation as in Equation 11:

$$Crash\,Risk_{i,t+1} = \alpha + \beta_0 Treated + \beta_1 Post + \beta_2 Treated \times Post + \gamma Controls_{i,t} + \epsilon_{i,t} \quad (11)$$

Where I use subscript $t$ because I use 12 months of data for both before and after periods to improve test power. Specifically, I run two sets of tests: first, I run the diff-in-diff test with cluster robust standard errors per Petersen (2009), clustering on both firm and time level. Second, I add firm and time fixed effects, which would absorb the treatment and post dummies, leaving the interaction term intact. The dependent variable is the estimated ex-ante monthly crash risk. $Treatment$ is a dummy variable that equals one if both firm ticker and option are mentioned in comments in Wallstreetbets in 2018, and zero otherwise. $Post$ is a dummy variable that equals one if the year is 2018, and zero otherwise. I also separately add controls to account for imperfect matching from possibly confounding factors. The results are reported in Table 11.

[Table 11 about here.]

As shown in table 11, the coefficient of interest is the interaction term, which accounts for the difference in treatment effect. The interaction term between $Treatment$ and $Post$ is significantly positive across all specifications, even after controlling for a battery of possible confounding firm characteristics. The estimated average effect is between around 1% to 1.6%, at less than 1% statistical significance level. This is strong evidence that retail participation tends to significantly increase stock ex-ante monthly crash risk.

The next important question is whether the effect of retail participation is stronger in smaller firms. As is often shown in literature, retail investors are more likely marginal price setters for smaller stocks, where arbitrage is costly (Pontiff, 1996). It follows naturally that in the case of ex-ante monthly crash risk, retail investors should have a greater impact on smaller firms. To test this hypothesis, I subset the firms at the beginning of 2017 into two groups, one with market value greater than the cross-sectional median, the other lower than the median. In this way, I generate a dummy variable $Big = 1$ if it belongs to the larger cohort, or zero otherwise. Then I conduct a triple difference-in-difference analysis, where I interact $Treatment$, $Post$, and $Big$ in the same setting as the prior tests, such that the

triple interaction term can be interpreted as the incremental treatment effect on large firms. The resulting specification can be represented as follows:

$$\begin{aligned} Crash\,Risk_{i,t+1} = \alpha &+ \beta_0 Treated + \beta_1 Post + \beta_2 Treated \times Post \\ &+ \beta_3 Big + \beta_4 Post \times Big + \beta_5 Treated \times Big \\ &+ \beta_6 Treated \times Post \times Big + \gamma Controls_{i,t} + \epsilon_{i,t} \end{aligned} \tag{12}$$

As before, I first run panel regressions with clustered standard errors on both firm and time level, and then run another test with firm and time fixed effects. The results are shown in Table 12.

[Table 12 about here.]

Table 12 shows evidence that, consistent with the literature, retail participation has a larger impact on the ex-ante monthly crash risk of smaller firms, while the impact on large firms is more muted on average. The coefficient on the interaction between $Treatment$ and $Post$ can be read as the effect on small firms, which is statistically significant and positive, which means that retail participation will on average increase the ex-ante crash risk of smaller than median size firms by about 1.4% to 1.9%. The coefficient on the triple interaction between $Treatment$, $Post$, and $Big$ is statistically significant and negative, which means that the retail impact on larger firms is smaller by about 0.6% to 1%.

The above results are done through examining all stocks that are available at the time in the sample with valid data. However, there might be legitimate concern that there is still underlying variables that correlate with being selected as treatment, that might confound the results. To alleviate that concern, I also perform propensity score matching before conducting the diff-in-diff analysis. Specifically, at the beginning of the sample (January 2017), I run a logistic regression of the dummy variable $Treatment \in 0, 1$ on the pertinent explanatory variables. These variables include: size, past three-month excess return, detrended turnover, total volatility, total skewness, asset growth, tangibility, sales growth, return on equity,

firm age, book-to-market ratio, SMIRK, relative deep out-of-money put option price, and relative deep out-of-money call option price. Then I generate the propensity score for each stock based on the fitted values of the logistic regression. For each treatment stock, I find the five stocks that have the closest propensity scores to the treatment stock, and randomly select two of them, with replacement. In this way, I match each treatment stock with at least one control stock. Then I run the same specifications as before. The results are presented in Table 13.

[Table 13 about here.]

Table 13 shows that, consistent with prior results using full sample, with PSM matched control firms, the treatment stocks display increased ex-ante crash risk by around 1% to 2.2%, depending on the specification. Moreover, there is consistent evidence that this effect differs between big and small firms: the effect on larger firms is around 0.6% to 1.7% less than the small firms, supporting the notion that retail investors might be marginal price setters for small firms.

Finally, it would also be interesting to see whether retail participation will impact the underlying variables that I use to predict ex-ante monthly crash probabilities. This would point to some channels that could also partially drive the increase of crash risk.[18] I choose the following dependent variables to examine: the relative deep out-of-money put and call option prices; trading volume as volume scaled by shares outstanding; total return volatility; and total return skewness. I follow the last test to run a triple difference-in-difference specification, with firm characteristics as controls. Therefore, the variables of interest are the interaction between $Treatment$ and $Post$, and the triple interaction between $Treatment$, $Post$, and $Big$. I present the results by using firm and time clustered standard errors adjustment in Table 14.[19]

[18]Note that in Table 11, I add predictor variables in the last two specifications as controls, and the results are still robust. But nonetheless it is interesting to examine the additional effects of retail participation on firm characteristics.

[19]I also ran panel regressions with firm and time fixed effects, and the results are largely similar. I omit them for brevity.

[Table 14 about here.]

Consistent with intuition, across the board, there is positive treatment effect for the five variables, as shown by the interaction term between $Treatment$ and $Post$, though the coefficients for trade volume and total return skewness are not significant. In addition, all the triple interactions between $Treatment$, $Post$, and $Big$ are shown as negative, supporting the prior finding that retail investors have a much less impact on bigger firms. One interesting results is that the relative prices of both deep out-of-money put and call options are significantly increased for small firms, suggesting a larger demand for these options, but the effect for large firms is muted since the triple interaction offsets it almost entirely. There is further anecdotal evidence that on Wallstreetbets, traders often boast how they trade options on small stocks. Another interesting results is that retail participation tends to significantly increase firm's stock return volatility, consistent with the findings in Foucault et al. (2011), and the effect is smaller for larger firms.

Taken together, these results enrich our understanding of how retail investors shape the tail risks of firms. Overall, the experiment provides evidence that retail participation would significantly increase firm ex-ante monthly crash risk, and a host of related underlying characteristics. Moreover, these effects are stronger in smaller firms, consistent with theory and empirical evidence. In other words, retail investors tend to make the left tail fatter, while chasing the left tail.

# 7. Robustness Tests

## *7.1. Alternative Simple Models*

In this section I present results from using alternative machine learning models with the same set of 15 variables as in the main results, where I use the same rolling window estimation procedure, but use three-fold cross validation to tune the model. The zero-cost portfolio alphas for these models are presented in Table 15.

[Table 15 about here.]

The results show that the estimates of ex-ante monthly crash risk are robust to different underlying estimating models.

### *7.2. More Complex Models*

One of the benefits of using machine learning models, more specifically, shrinking methods, is the ability to utilize more independent variables, or "features". In other words, we can use more conditioning information. In extreme situations, the number of features can be greater than the number of observations, while the rank condition in such situation dictates that ordinary least squares (OLS) would not have unique solutions. To explore this advantage, I enlarge the set of features from 15 variables to 134 variables. On top of the original variables, I add past one-month excess return, past year accruals per Dechow et al. (1995) (Modified Jones Model), and option-based variables. To take advantage of the rich information from option market, I create option-related variables as follows: for each stock each month, I divide the options into calls and puts; for calls, they are categorized further into four groups per moneyness: $[1.05, 1.1]$, $(1.1, 1.15]$, $(1.15, 1.2]$, and $(1.2, \infty)$; likewise, for puts, they are grouped into four moneyness classes: $(0, 0.8]$, $(0.8, 0.85]$, $(0.85, 0.9]$, and $(0.9, 0.95]$; finally, compute the mean, median, min, max, and standard deviation for relative option price (as defined in previous sections), volume, and implied volatility. Therefore, there are in total $2 \times 4 \times 5 \times 3 = 120$ option-related variables.

To handle large set of features efficiently, and to increase their signal-to-noise ratio, I apply principal component transformation (PCA) to these variables to reduce dimensionality before feeding them into machine learning models.[20] The procedure is as follows: for each window, first standardize the training set, and use the information to transform validation data and test data; second, transform the resulting training set into either 5, 10, or 20 principal

[20]This is to ensure that the computation can be done within a short period of time, mimicking the urgency of trading by investors; otherwise, the computation can be unwieldy and takes too much time, which might be unrealistic.

components, and use the underlying transformation to transform the validation and test data; finally, apply SMOTE to balance the training data. With the transformed data ready, three machine learning models are used: ridge regression; XGBoost tree methods (Chen and Guestrin, 2016); and feed forward neural network (multi-layer perceptrons, or MLP).

I show in Appendix, that using more complex models produce quantitatively similar results. Therefore, to maintain parsimony and interpretability, I present simple model as the main results.

### *7.3. Endogeneity Concern for the Experiment*

There is concern for possible endogeneity for the quasi-natural experiment. Due to data limitation, we cannot perfectly observe whether the traders post their positions on "Wallstreetbets" after their trade, or simply trade whatever they see on Wallstreetbets. For this I offer one possible counter argument. That is, the two flairs I used, "Daily discussions" and "What's your move tomorrow?", only started at the end of 2017, coinciding with the timing that Robinhood announced they would offer commission-free trading. Before that, there were no regular posts, but only random individual posts scattered on Wallstreetbets. Therefore it would be much less likely that traders would scout over Wallstreetbets for stock actions and trade accordingly. Even if we accept the assumption that there are indeed a subset of Robinhood traders would participate in such a trade, what we would capture in our setting is the incremental demand from traders that emerged after the introduction of commission-free option trading, which jointly represent the demand for stocks and their underlying options.

Second, if indeed there is influx of option traders on Robinhood as we hypothesized, we should observe that more Reddit users would post comments on "Wallstreetbets" concerning Robinhood. This should be another piece of evidence that the introduction of commission-free option trading induced more traders to jump in. To see this, I scraped all posts on "Wallstreetbets" that contain keywords "Robinhood", "robinhood", "ROBINHOOD" or "RH",

as the search keywords are case sensitive. Then I count the number of posts per day and number of authors that made these comments per day, and plot them in Figure 8.

[Fig. 8 about here.]

As shown in Figure 8, the blue line is number of posts, while the red line is number of unique authors/users. The green vertical line indicates the last day of 2017 as the watershed of the event where Robinhood introduced commission-free option trading. Between the beginning of 2017 and the end of 2018, there are in total 2913 posts and 2784 users on "Wallstreetbets" that mention Robinhood keywords. Before the end of 2017, both the number of posts and number of authors that mentioned "Robinhood" remained quite stable. The mean daily number of posts before the end of 2017 is approximately 2.7, and the mean daily number of users that mention Robinhood keywords before the end of 2017 is approximately 2.6. After the event, mean daily number of posts becomes 6.3, and mean daily number of users becomes 6.0, more than double those before the event. Moreover, a simple $t$-test shows that the differences in means before and after the event are highly statistically significant, with $p$-values well below 0.001. Therefore, this provides evidence that there was indeed significant increases in users that paid more attention to Robinhood after the introduction event, and consequently reinforce our claim that the event could have potentially induced more trading on Robinhood.

To further bolster this claim, I conduct a separate test to see whether there is significant increase in option trading volume after the introduction of commission-free trading. If Robinhood users began to trade options because they see comments from "Wallstreetbets", then they should be equally likely to do so before and after the introduction of commission-free trading. Thus we should not see a significant increase in option trading volume around the event. If instead there is significant increase in option trading volume around the event, after controlling for other factors, then the event should be a major factor that causes the increase of option trading volume, thus validating the proposed experiment. I first plot the daily total volume of out-of-money option trading before and after the event in Figure 9.

[Fig. 9 about here.]

As shown in Figure 9, there is dramatic increase in the out-of-money option trading volume for both puts and calls after the end of 2017. Interestingly, there is a significant jump in volume during the first three months of 2018. This could be correlated with the rolling out of option trading feature. However without further evidence, this is where we stop speculating. Nevertheless, comparing 2017 and 2018, there is an apparent increase in trading volume for out-of-money options.

Next I conduct formal analysis of whether there is significant increase in out-of-money option trading volume after the event, especially for those treatment stocks where I outlined the identification strategy in the previous section. Control groups are selected via PSM matching, which is also outlined in the previous section. The results are shown in Table 16.

[Table 16 about here.]

In Table 16, Panel A regresses log volume of out-of-money daily put options on treatment, post, and big dummies, their interactions, and control variables, while Panel B examines calls. The tests use either two-way clustered standard errors or fixed effects. As shown in both panels, the results conform nicely with the difference-in-difference tests we have shown in the previous section. That is, there is significant increase in total trading volume of both out-of-money puts and calls after the introduction of commission-free option trading. Moreover, the increase is more dramatic in small stock options. This is an important piece of evidence that the experiment did provide a shock to the demand for out-of-money option trading, thus supporting our use of this shock as a quasi-natural experiment.

# 8. Conclusion

Tails are fat, and they are fatter now thanks to higher retail participation in the stock market. This paper shows that a subset of retail investors, proxied by "Robinhood Traders"

and "Reddit Traders", through their chasing of extreme returns, also known as lottery characteristics, reinforce this characteristic.

This study builds on prior literature to develop an ex-ante measure for firm-level monthly crash and jackpot probabilities via machine learning. I combine novel imbalanced learning techniques with Ridge regression to show superior forecasting power for subsequent one month crashes and jackpots. The estimated crash risk is robustly priced both in time-series portfolio tests and cross-sectional tests. High crash risk stocks are difficult to short, since they tend to be smaller and less liquid. Building on these findings, I show that retail investors, proxied by Robinhood traders, seem to buy the left tail, likely causing high crash risk stocks overpriced, which subsequently leads to lower returns. Using Robinhood introduction of commission-free option trading at the end of 2017 as a quasi-experiment setting, together with textual information from Reddit, I show that retail participation significantly increased ex ante stock crash risk, and this effect is stronger for small firms.

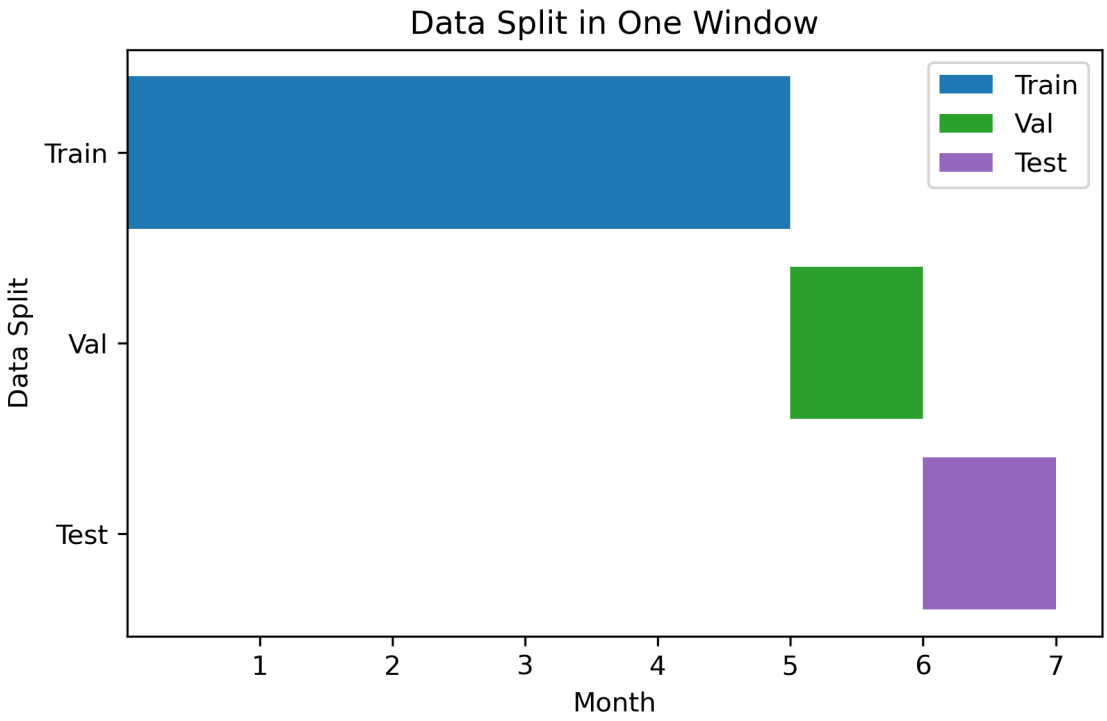


Fig. 1. Data Split in Window. This figure plots how the data of one rolling window is split. The bars represent months of data in a window. From top to bottom: the first bar represents the training set, which consists of five months of data; the second bar is the validation set, consisting one month of data; the last bar is the test set, consisting one month of data. For example, the first rolling window consists training data (January 1996 to May 1996), validation data (June 1996), and test data (July 1996). The training set is used to fit the model; the validation set is used to tune the hyper parameters to find the best estimator in terms of forecasting metric (e.g., F1 score); the resulting optimal estimator is then used to fit the test data, and compare the prediction with the ground truth, in order to generate the test performance metrics.

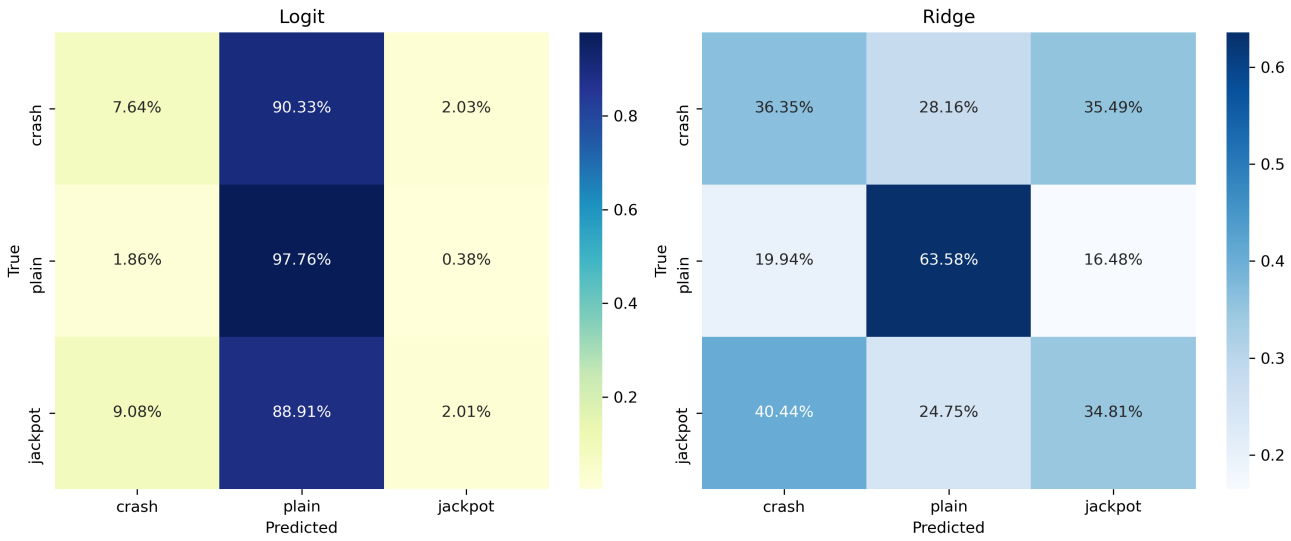


Fig. 2. Aggregate Confusion Matrices. This figure plots the aggregate confusion matrices for simple logit and Ridge, where the predicted classes are add up across time. The rows are true classes, while the columns are predicted classes. All elements are normalized row by row, such that the diagonal elements can be viewed as recall for each class.

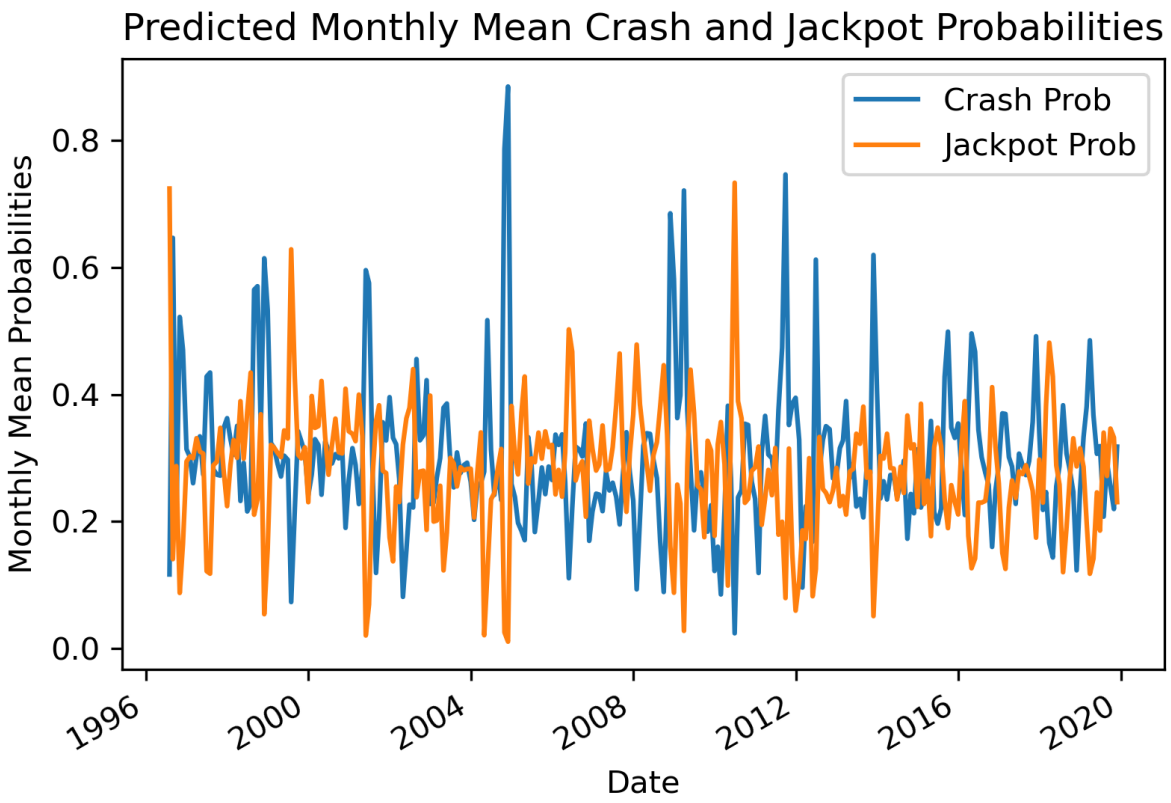


Fig. 3. Mean Monthly Predicted Crash and Jackpot Probabilities. This figure plots the mean monthly predicted crash and jackpot probabilities over time, per Ridge model. Each month, I calculate the cross-sectional mean predicted crash and jackpot probabilities respectively, and then plot them against time. The sample runs from July 1996 to December 2019.

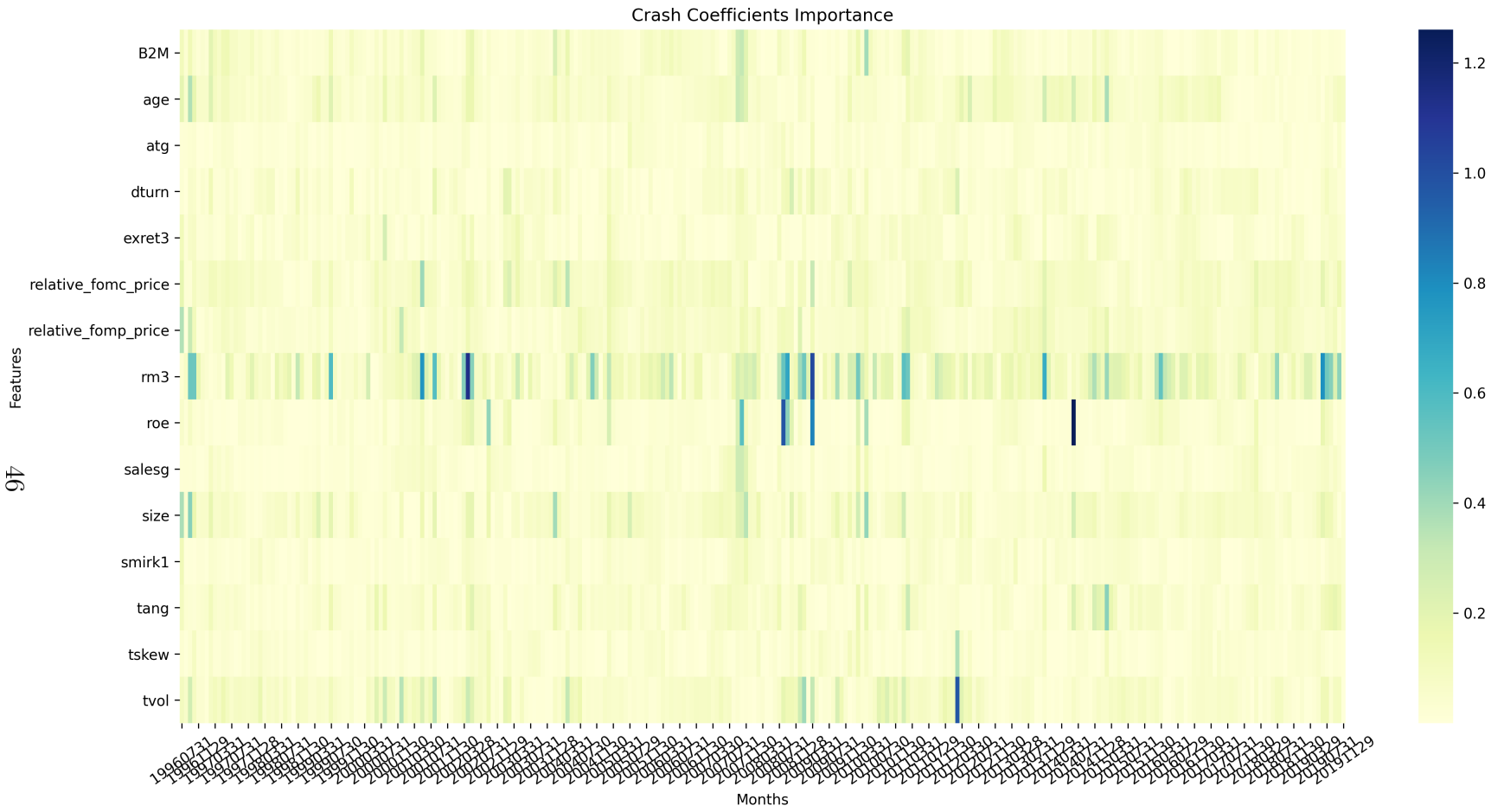


Fig. 4. Crash Coefficients Importance. This figure plots the absolute values of each coefficient for crash class for each of the 281 estimation window. All variables are standardized before training, and thus the absolute values represent relative "size" of their importance. The sample runs from July 1996 to December 2019.

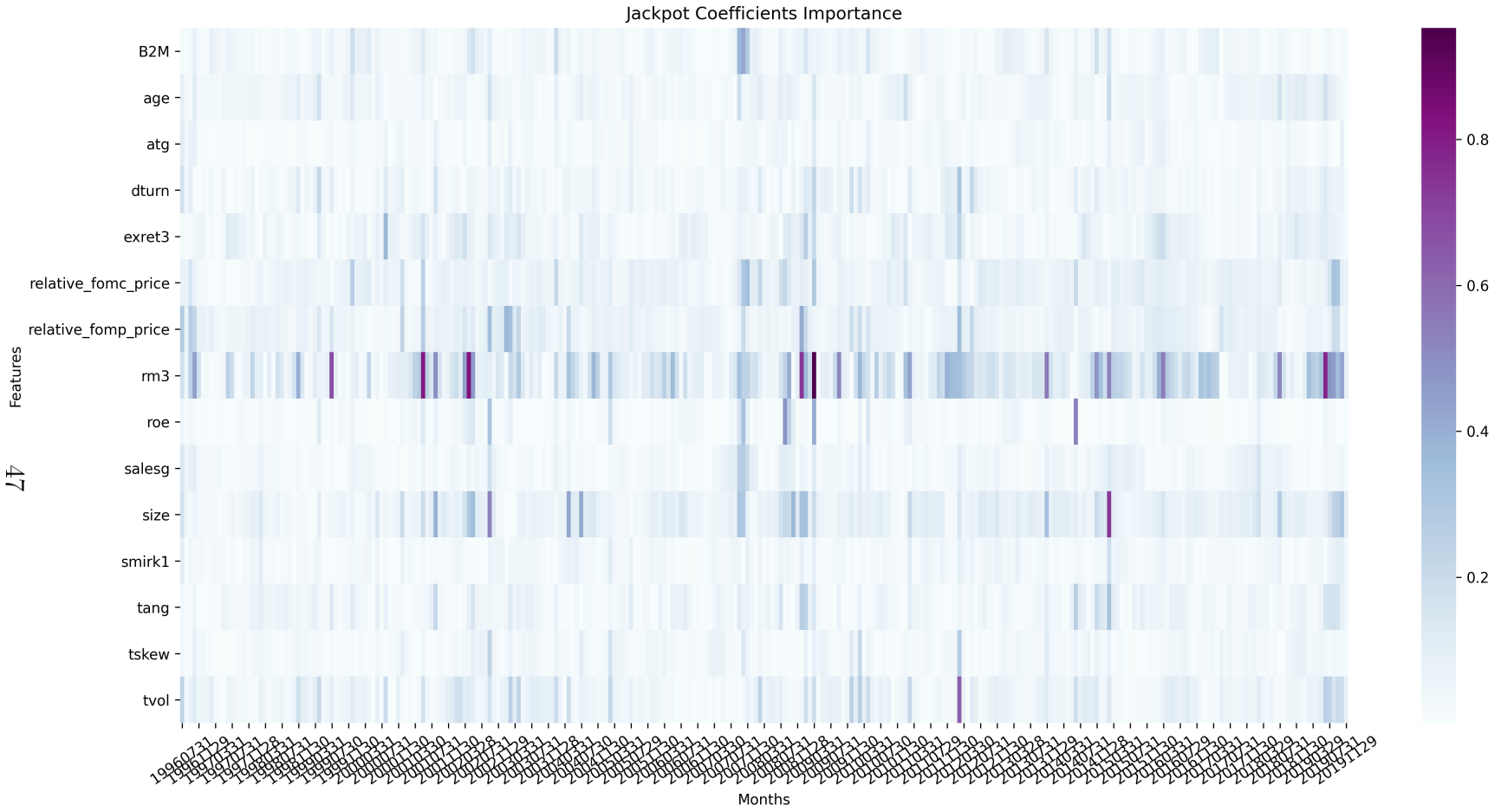


Fig. 5. Jackpot Coefficients Importance. This figure plots the absolute values of each coefficient for jackpot class for each of the 281 estimation window. All variables are standardized before training, and thus the absolute values represent relative "size" of their importance. The sample runs from July 1996 to December 2019.

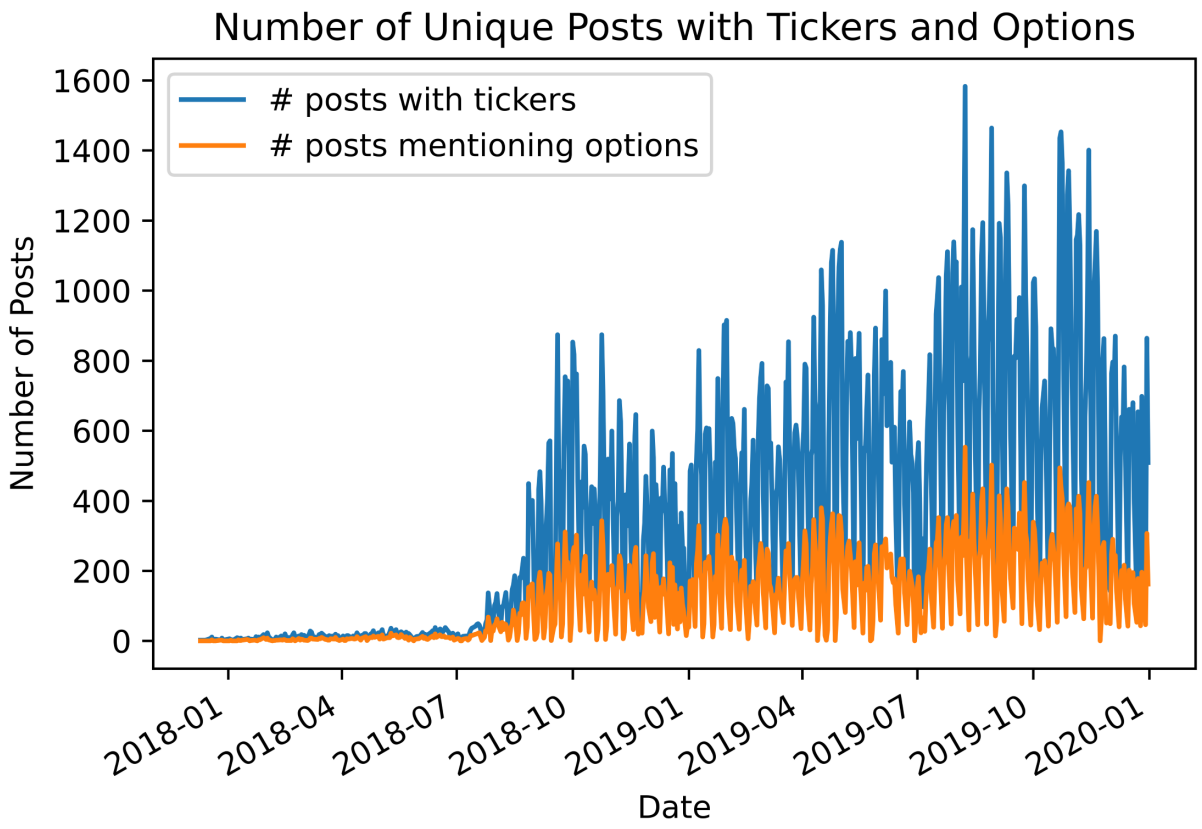


Fig. 6. Number of Posts Over Time. This figure plots the number of unique posts that contain ticker names, and of which, number of posts that mention options on Wallstreetbets of Reddit. The sample runs from December 2017 to December 2019.

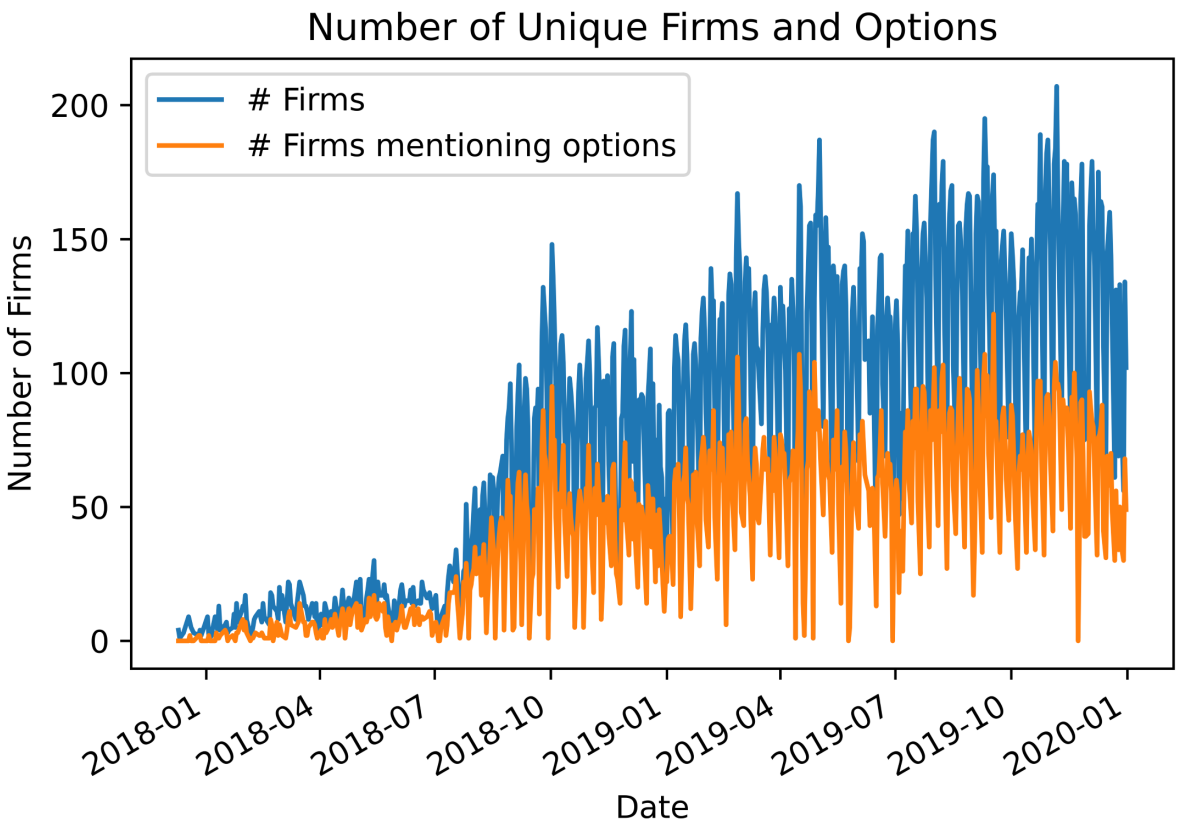


Fig. 7. Number of Firms Mentioned Over Time. This figure plots the number of unique firms that were mentioned, and of which, number of firms that are also co-mentioned with options on Wallstreetbets of Reddit. The sample runs from December 2017 to December 2019.

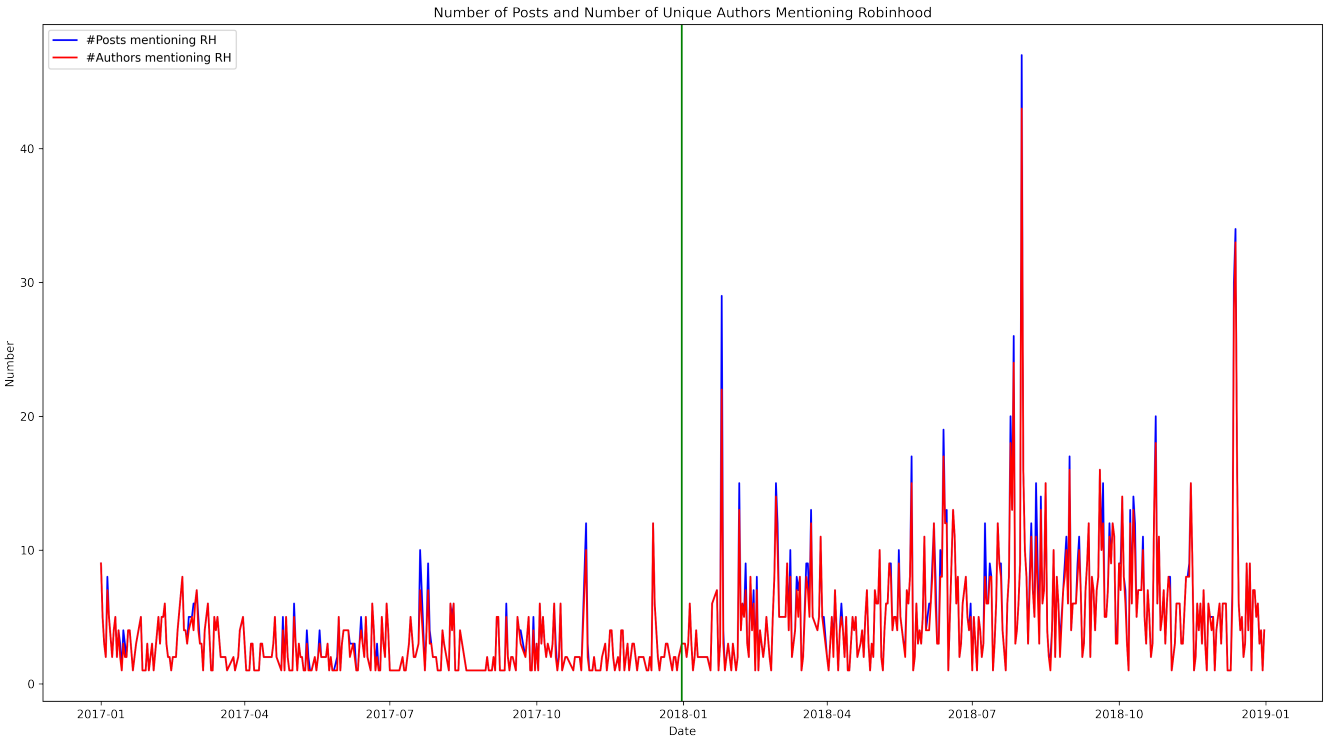


Fig. 8. Number of Posts and Number of Authors Mentioning Robinhood. This figure plots the number of posts and number of authors that mentioned the keywords "Robinhood", "robinhood", "ROBINHOOD" or "RH" on "Wallstreetbets", the subreddit. The sample runs from January 2017 to December 2018. Numbers are displayed in daily frequency. The blue line is number of posts, while the red line is number of unique authors/users. The green vertical line indicates the last day of 2017 as the watershed of the event where Robinhood introduced commission-free option trading.

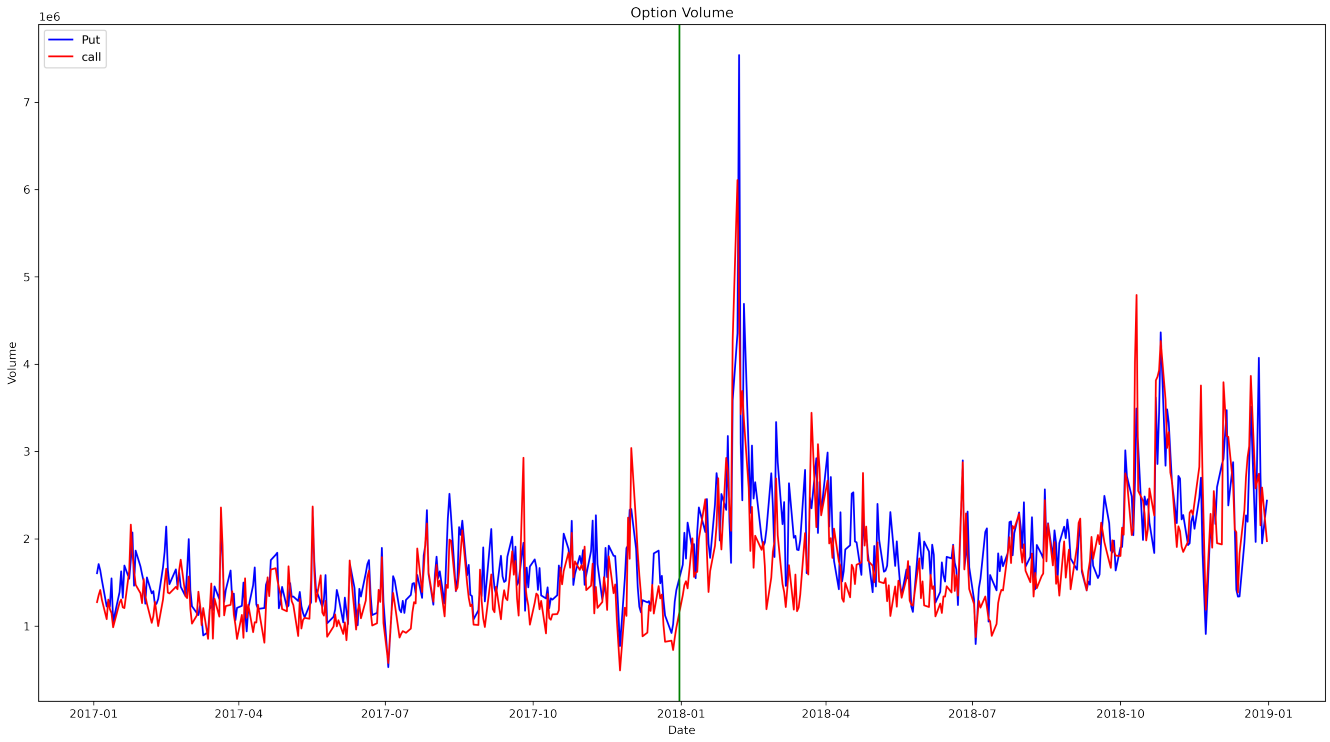


Fig. 9. Daily Volume of Out-of-Money Option Trading. This figure plots the daily total volume of out-of-money option trading. The blue line is put volume, and red line call volume. Out-of-money puts are defined as moneyness less than or equal to 0.95, while out-of-money calls are defined as moneyness greater than or equal to 1.05. For both puts and calls, the following filter is used: days to expiration between 10 and 180 days, implied volatility between 0.03 and 2, positive open interest, option price greater than or equal to \$0.125, and non-missing volume. Numbers are displayed in daily frequency. The green vertical line indicates the last day of 2017 as the watershed of the event where Robinhood introduced commission-free option trading. The sample starts from the beginning of 2017 to the end of 2018.

Table 1: Summary Statistics

The table presents summary statistics for key variables used in my baseline model for predicting crash probabilities. For sake of brevity, I present characteristics for crashes and jackpots, omitting "plain" cases. The total observations amount to 403,379 firm×month pairs, with "plain" cases at 358,780 observations. I define crashes as monthly log return of less than -20%, and jackpots as monthly log return of greater than 20%. SMIRK is the implied volatility smirk measure per Xing et al. (2010). FOMP and FOMC are deep out-of-money put option and call option relative price measure per Barro and Liao (2020). Variable definitions can be found in Appendix. The sample runs from January 1996 to December 2019. All variables are lagged for one period.

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | crashes | | | | Plain | | | | jackpots | | | |
| VARIABLES | mean | sd | min | max | mean | sd | min | max | mean | sd | min | max |
| Size | 20.74 | 1.340 | 16.73 | 27.60 | 21.64 | 1.570 | 16.45 | 27.80 | 20.60 | 1.281 | 16.66 | 27.06 |
| Exret3 | -0.033 | 0.322 | -2.295 | 2.436 | 0.007 | 0.193 | -2.235 | 2.449 | -0.038 | 0.340 | -2.142 | 2.377 |
| RM3 | -0.004 | 0.096 | -0.396 | 0.253 | 0.025 | 0.076 | -0.396 | 0.253 | 0.004 | 0.100 | -0.396 | 0.253 |
| Dturn | -0.002 | 0.300 | -10.42 | 10.16 | 0.0002 | 0.145 | -7.975 | 11.31 | -0.007 | 0.265 | -6.073 | 13.39 |
| Tvol | 0.042 | 0.025 | 0.007 | 0.411 | 0.025 | 0.016 | 0.001 | 0.440 | 0.041 | 0.023 | 0.004 | 0.333 |
| Tskew | 0.081 | 0.935 | -4.644 | 4.626 | 0.072 | 1.020 | -4.673 | 4.686 | 0.106 | 0.945 | -4.542 | 4.574 |
| ATG | 0.043 | 0.201 | -2.968 | 3.029 | 0.0306 | 0.126 | -2.968 | 3.723 | 0.0403 | 0.183 | -1.817 | 3.723 |
| Tang | 0.315 | 0.422 | 0 | 5.869 | 0.351 | 0.428 | 0 | 7.933 | 0.322 | 0.413 | 0 | 6.268 |
| Salesg | 0.167 | 0.650 | -7.857 | 9.270 | 0.110 | 0.403 | -9.901 | 12.43 | 0.172 | 0.605 | -9.901 | 7.496 |
| ROE | -0.067 | 44.02 | -5,958 | 3,162 | 0.026 | 16.35 | -5,958 | 3,162 | 0.049 | 5.554 | -49.72 | 534.3 |
| Age | 13.81 | 14.42 | 0 | 93 | 21.83 | 19.48 | 0 | 93 | 13.85 | 14.07 | 0 | 93 |
| B2M | 0.540 | 0.649 | 0 | 18.57 | 0.476 | 0.444 | 0 | 25.85 | 0.588 | 0.780 | 0.0002 | 32.77 |
| SMIRK | 0.068 | 0.078 | -1.056 | 1.355 | 0.053 | 0.050 | -1.093 | 1.433 | 0.065 | 0.069 | -0.558 | 1.069 |
| FOMP_price | 0.040 | 0.025 | 0.002 | 0.248 | 0.022 | 0.017 | 0.001 | 0.239 | 0.041 | 0.025 | 0.002 | 0.253 |
| FOMP_price | 0.047 | 0.026 | 0.002 | 0.265 | 0.028 | 0.019 | 0.001 | 0.339 | 0.048 | 0.026 | 0.003 | 0.298 |
| Obs | 25,355 | | | | 358,780 | | | | 19,244 | | | |

Table 2: Mutlinomial Logit
The table runs a multinomial logit regression predicting crashes and jackpots for sample period 1996 - 2019. "plain" cases are set as base and are omitted. Variable definitions are shown in Appendix. Each variable is properly lagged. The crashes and jackpots are classified as one-month ahead monthly log returns of less than -20% and greater than 20% respectively. SMIRK is the implied volatility smirk measure per Xing et al. (2010). FOMP and FOMC are deep out-of-money put option and call option relative price measure per Barro and Liao (2020), and is computed as option price scaled by forward stock price. All independent variables are standardized to aid interpretation and comparison. Standard errors are clustered at stock and month levels per Petersen (2009) and are included in parentheses.

| | (1)<br>Crash | (2)<br>Jackpot |
|---|---|---|
| Relative_FOMP_price | 0.149*** | 0.121*** |
| | (0.022) | (0.027) |
| Relative_FOMC_price | 0.264*** | 0.357*** |
| | (0.031) | (0.030) |
| SMIRK | -0.008 | -0.032 |
| | (0.018) | (0.020) |
| RM3 | -0.170** | -0.110* |
| | (0.069) | (0.063) |
| Exret3 | -0.022 | -0.054* |
| | (0.022) | (0.029) |
| B2M | -0.046 | 0.040 |
| | (0.033) | (0.028) |
| ATG | 0.036*** | 0.025* |
| | (0.010) | (0.013) |
| ROE | -0.076*** | 0.040*** |
| | (0.013) | (0.015) |
| Tvol | 0.562*** | 0.490*** |
| | (0.057) | (0.052) |
| Tskew | -0.006 | 0.011 |
| | (0.010) | (0.012) |
| Size | -0.207*** | -0.376*** |
| | (0.048) | (0.045) |
| Dturn | -0.115*** | -0.110*** |
| | (0.019) | (0.016) |
| Age | -0.173*** | -0.134*** |
| | (0.022) | (0.023) |
| Tang | 0.028 | 0.039** |
| | (0.023) | (0.019) |
| Salesg | 0.044*** | 0.066*** |
| | (0.015) | (0.016) |
| Observations | 403,379 | |
| Pseudo R2 | 0.123 | |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 3: Mean Performance Metrics

The table reports mean performance metrics for simple logit and Ridge across the rolling prediction windows from January 1996 to December 2019. Each window consists of 6-month training set and 1-month test set. In the case of simple logit, the whole training set is fitted and used to fit the test set. In the case of Ridge, the training set is further split into 5 months of training data and 1 month of validation data, where the training data is used to tune the Ridge estimator (through penalty factor $\lambda$), and then the best estimator is chosen to fit the test set. The metrics are defined as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$
$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$
$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These metrics are computed for each of the three classes. There are in total 281 windows, and hence 281 sets of metrics are generated in total for each underlying model. These metrics are then averaged across time.

| Class | Metrics | logit | Ridge |
|---|---|---|---|
| Crash | Precision | 0.177 | 0.128 |
| | Recall | 0.062 | 0.412 |
| | F1 | 0.049 | 0.128 |
| Plain | Precision | 0.891 | 0.935 |
| | Recall | 0.970 | 0.626 |
| | F1 | 0.922 | 0.730 |
| Jackpot | Precision | 0.100 | 0.090 |
| | Recall | 0.014 | 0.344 |
| | F1 | 0.018 | 0.108 |

Table 4: Decile High-Minus-Low Alphas
This table presents the high-minus-low long-short zero-cost strategy alphas, per asset pricing model, for both equal-weighted and value-weighted portfolios. At the end of each month, stocks are ranked by their ex-ante crash probabilities produced by Ridge model into ten decile portfolios each month. Then the high-minus-low return series for both equal-weighted and value-weighted returns where we long highest decile portfolio and short lowest decile portfolio, are regressed on various risk factor return series. The asset pricing models include: CAPM market model, Fama-French three-factor model (FF3) (Fama and French, 1993), then augmented with a momentum factor (FF4) (Carhart, 1997), Fama-French five-factor model (FF5) (Fama and French, 2015), and finally FF5 augmented with momentum factor (FF6). $t$-statistics are included. Time-series regressions are estimated with Newey-West standard errors with 12 lags.

| | Value-weighted | | Equal-weighted | |
|---|---|---|---|---|
| Pricing_model | Alpha | $t$-stat | Alpha | $t$-stat |
| CAPM | -1.523*** | -2.999 | -1.594*** | -3.303 |
| FF3 | -1.467*** | -3.810 | -1.557*** | -4.259 |
| FF4 | -1.054*** | -2.861 | -1.125*** | -3.272 |
| FF5 | -0.932*** | -2.881 | -1.186*** | -3.567 |
| FF6 | -0.664** | -2.384 | -0.898*** | -3.289 |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 5: Value-Weighted Decile Portfolio Regressions

Stocks are ranked by their ex-ante crash probabilities each month into ten decile portfolios. Panel A presents the alphas from regressing the excess returns of the ten value-weighted decile portfolios, per each asset pricing model. The asset pricing models include: CAPM market model, Fama-French three-factor model (FF3) (Fama and French, 1993), then augmented with a momentum factor (FF4) (Carhart, 1997), Fama-French five-factor model (FF5) (Fama and French, 2015). Panel B reports the coefficients from regressing the excess returns of the decile portfolios on FF5 factors augmented with momentum factor (FF6). Newey-West standard errors with 12 lags are included in parentheses.

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Panel A: Portfolio Alphas** | | | | | | | | | | |
| CAPM | 0.142* | 0.215** | 0.157 | 0.163 | -0.179 | -0.310** | -0.454** | -0.652** | -0.659* | -1.381*** |
| | (0.083) | (0.086) | (0.115) | (0.129) | (0.111) | (0.135) | (0.180) | (0.269) | (0.377) | (0.440) |
| FF3 | 0.136** | 0.193** | 0.124 | 0.151 | -0.193** | -0.315** | -0.445*** | -0.621*** | -0.653** | -1.331*** |
| | (0.062) | (0.076) | (0.097) | (0.119) | (0.090) | (0.124) | (0.142) | (0.203) | (0.311) | (0.344) |
| FF4 | 0.049 | 0.151* | 0.078 | 0.181 | -0.158* | -0.223* | -0.278** | -0.453** | -0.361 | -1.005*** |
| | (0.063) | (0.089) | (0.103) | (0.121) | (0.084) | (0.117) | (0.120) | (0.200) | (0.293) | (0.336) |
| FF5 | -0.011 | 0.062 | 0.057 | 0.096 | -0.173** | -0.298** | -0.265 | -0.427*** | -0.346 | -0.944*** |
| | (0.068) | (0.083) | (0.103) | (0.117) | (0.087) | (0.132) | (0.160) | (0.157) | (0.283) | (0.300) |
| **Panel B: FF6 Factor Coefficients** | | | | | | | | | | |
| Intercept | -0.066 | 0.038 | 0.028 | 0.120 | -0.151* | -0.232* | -0.158 | -0.315** | -0.152 | -0.729*** |
| | (0.066) | (0.092) | (0.105) | (0.120) | (0.084) | (0.125) | (0.111) | (0.158) | (0.235) | (0.275) |
| Mkt_RF | 0.887*** | 1.036*** | 1.141*** | 1.157*** | 1.184*** | 1.241*** | 1.273*** | 1.314*** | 1.298*** | 1.491*** |
| | (0.030) | (0.035) | (0.026) | (0.042) | (0.038) | (0.041) | (0.064) | (0.096) | (0.128) | (0.141) |
| SMB | -0.160*** | -0.046* | 0.054 | 0.139** | 0.273*** | 0.294*** | 0.269** | 0.379** | 0.439** | 0.396* |
| | (0.042) | (0.028) | (0.044) | (0.058) | (0.067) | (0.075) | (0.107) | (0.165) | (0.187) | (0.212) |
| HML | -0.003 | 0.084** | 0.154** | 0.023 | 0.044 | -0.129 | -0.084 | -0.270** | -0.218* | -0.424** |
| | (0.035) | (0.034) | (0.068) | (0.059) | (0.102) | (0.101) | (0.075) | (0.117) | (0.112) | (0.166) |
| RMW | 0.139* | 0.206*** | 0.088* | 0.192*** | 0.080 | 0.015 | -0.142 | -0.224 | -0.425* | -0.541** |
| | (0.073) | (0.058) | (0.046) | (0.052) | (0.070) | (0.121) | (0.146) | (0.179) | (0.242) | (0.261) |
| CMA | 0.231** | 0.082 | 0.045 | -0.122 | -0.199* | 0.020 | -0.258 | -0.151 | -0.060 | -0.119 |
| | (0.091) | (0.060) | (0.089) | (0.087) | (0.112) | (0.128) | (0.159) | (0.157) | (0.156) | (0.189) |
| UMD | 0.115*** | 0.050 | 0.062 | -0.049 | -0.048** | -0.135*** | -0.230*** | -0.232*** | -0.408*** | -0.450*** |
| | (0.022) | (0.031) | (0.051) | (0.031) | (0.021) | (0.051) | (0.050) | (0.051) | (0.073) | (0.103) |
| R-squared | 0.854 | 0.901 | 0.892 | 0.871 | 0.882 | 0.873 | 0.842 | 0.822 | 0.788 | 0.779 |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 6: FMB Cross-Sectional Regressions
This table reports Fama-MacBeth regressions of raw returns on lagged firm characteristics in the spirit of Fama and French (2020). Independent variables are standardized cross-sectionally each month. Control variables include : size, book-to-market ratio, asset growth, ROE, momentum, short-term reversal. In Column (3), I add MAX (Bali et al., 2011), which is the highest daily return of the past month. In column (4), I add illiquidity $ILLIQ$ (Amihud, 2002), turnover $TURN$, and idiosyncratic volatility $IVOL$. In Column (5), I add other anomaly variables: abnormal capital investment $ACI$ (Titman et al., 2004), asset growth $AG$ (Cooper et al., 2008), composite equity issues $CEI$ (Daniel and Titman, 2006), gross profitability $GP$ (Novy-Marx, 2013), net operating assets $NOA$ (Hirshleifer et al., 2004), net stock issues $NSI$ (Ritter, 1991), and O-score $OSCR$ (Ohlson, 1980). Standard errors are adjusted according to Newey-West procedures.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Dep Var: Returns | | | | |
| Crash_prob | -0.274*** | -0.258*** | -0.267*** | -0.251*** | -0.308*** |
| | (0.065) | (0.084) | (0.089) | (0.088) | (0.098) |
| Jackpot_prob | 0.352*** | 0.369** | 0.361* | 0.383* | 0.429* |
| | (0.097) | (0.172) | (0.200) | (0.219) | (0.259) |
| Size | | -0.151 | -0.144 | -0.059 | -0.064 |
| | | (0.103) | (0.112) | (0.131) | (0.149) |
| B2M | | -0.107** | -0.092** | -0.067* | 0.009 |
| | | (0.048) | (0.043) | (0.040) | (0.069) |
| ROE | | 0.315*** | 0.349*** | 0.354*** | 0.352*** |
| | | (0.036) | (0.049) | (0.051) | (0.040) |
| ATG | | -0.033 | -0.065 | -0.065 | -0.064 |
| | | (0.074) | (0.053) | (0.050) | (0.049) |
| REV | | -0.156*** | -0.157** | -0.251*** | -0.269*** |
| | | (0.048) | (0.068) | (0.081) | (0.089) |
| MOM | | 0.067 | 0.049 | 0.048 | 0.058 |
| | | (0.044) | (0.050) | (0.050) | (0.044) |
| MAX | | | 0.096 | 0.288** | 0.282** |
| | | | (0.070) | (0.112) | (0.115) |
| ILLIQ | | | | 2.010 | 1.747 |
| | | | | (2.071) | (1.893) |
| Turnover | | | | -0.032 | -0.043 |
| | | | | (0.061) | (0.073) |
| IVOL | | | | -0.240** | -0.288*** |
| | | | | (0.095) | (0.103) |
| Anomalies | NO | NO | NO | NO | YES |
| Observations | 398,604 | 398,604 | 398,604 | 398,604 | 398,604 |
| Average R2 | 0.010 | 0.031 | 0.035 | 0.050 | 0.069 |
| Number of groups | 281 | 281 | 281 | 281 | 281 |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 7: Value-Weighted Decile Portfolio Characteristics
Stocks are ranked by their ex-ante crash probabilities each month into ten decile portfolios. The table reports the mean characteristics of each decile portfolio. The characteristics include size, MAX, idiosyncratic volatility, and illiquidity. For each portfolio, each characteristic is averaged across member stocks and across time. Variable definitions are included in Appendix.

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Size | 23.669 | 22.756 | 22.256 | 21.862 | 21.530 | 21.222 | 20.954 | 20.689 | 20.407 | 20.044 |
| MAX | 3.754 | 4.050 | 4.437 | 4.811 | 5.196 | 5.650 | 6.242 | 6.914 | 7.737 | 10.072 |
| Ivol | 1.265 | 1.369 | 1.501 | 1.645 | 1.788 | 1.961 | 2.167 | 2.395 | 2.725 | 3.529 |
| Illiq | 0.011 | 0.013 | 0.019 | 0.025 | 0.033 | 0.042 | 0.056 | 0.070 | 0.093 | 0.147 |

Table 8: Short Interest and Crash Risk
This table reports panel regressions of monthly percentage of short interest on ex-ante crash risk, jackpot risk, and a host of other stock characteristics. All variables are standardized cross-sectionally to be mean zero and standard deviation of one to aid interpretability. Column (2) includes an interaction term between crash risk and lagged size to examine the differential effect between large and small firms. Robust standard errors are reported in parentheses.

| | (1) | (2) |
|---|---|---|
| VARIABLES | Dep Var: short Interest | |
| Crash_prob | 0.003 | 0.020*** |
| | (0.002) | (0.002) |
| Jackpot_prob | -0.007*** | -0.014*** |
| | (0.002) | (0.002) |
| Crash_prob×Size | | 0.045*** |
| | | (0.002) |
| Size | -0.630*** | -0.634*** |
| | (0.006) | (0.006) |
| B2M | -0.020*** | -0.016*** |
| | (0.003) | (0.003) |
| ROE | 0.000 | 0.000 |
| | (0.002) | (0.002) |
| ATG | -0.004*** | -0.005*** |
| | (0.001) | (0.001) |
| Lag_ret | 0.015*** | 0.015*** |
| | (0.001) | (0.001) |
| Turnover | 0.432*** | 0.428*** |
| | (0.003) | (0.003) |
| Tvol | -0.151*** | -0.149*** |
| | (0.002) | (0.002) |
| Tskew | 0.001 | 0.001 |
| | (0.001) | (0.001) |
| Tang | 0.010*** | 0.011*** |
| | (0.003) | (0.003) |
| Salesg | 0.007*** | 0.006*** |
| | (0.001) | (0.001) |
| Observations | 342,419 | 342,419 |
| R-squared | 0.614 | 0.615 |
| Firm & Time FE | YES | YES |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 9: Retail Trading Imbalance and Monthly Crash Risk
This table shows the results that examine the relationship between retail trading imbalance and monthly crash risk. In Column (1), I run Fama-MacBeth cross-sectional regressions to estimate the average coefficients on crash and jackpot risks, controlling for other firm characteristics, including $MAX$ measure as another proxy for lottery characteristics. In Column (2), I run panel regression, with both firm and time fixed effects to control for unobserved heterogeneities. Retail trading imbalance is defined as:

$$Change\#User = \log(\#User_{i,t}) - \log(\#User_{i,t-1})$$

The user data is from Robintrack, which provides hourly data on the number of users that hold a particular stock. The user change here is defined at monthly frequency. All variables are at [0.5%, 99.5%] level to remove the effects of outliers. The sample runs from June 2018 to December 2019 at monthly frequency.

| | (1) | (2) |
|---|---|---|
| | Dep Var: Retail%Imbalance | |
| VARIABLES | FMB | Panel |
| Crash_prob | 0.127** | 0.152*** |
| | (0.052) | (0.026) |
| Jackpot_prob | 0.311** | 0.222*** |
| | (0.123) | (0.025) |
| MAX | 0.773*** | 0.852*** |
| | (0.072) | (0.048) |
| Size | 0.014*** | 0.051*** |
| | (0.003) | (0.007) |
| B2M | 0.004 | 0.002 |
| | (0.004) | (0.008) |
| ROE | 0.002 | 0.007 |
| | (0.006) | (0.010) |
| ATG | 0.033*** | 0.009 |
| | (0.010) | (0.008) |
| Exret3 | 0.033*** | 0.025*** |
| | (0.008) | (0.006) |
| Ivol | -0.009 | -0.003 |
| | (0.006) | (0.003) |
| Tvol | -0.018** | -0.022*** |
| | (0.007) | (0.003) |
| FF3 $\beta$s | YES | YES |
| Observations | 27,159 | 27,105 |
| R-squared | 0.113 | 0.162 |
| Firm & Time FE | NO | YES |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 10: WSB Mentioning and Retail Trading
The table examines the relationship between Wallstreetbets co-mentioning of stock tickers and option keywords and retail trading for sample period between May 2018 to December 2019. Variables are measured at daily frequency. Dependent variables include: Robinhood trade imbalance, measured by the change in log of number of users; log of number of users, as a proxy for the popularity of a certain stock; trading volume; and change in trading volume. The control variables are all one-day lagged and include: price, log of Market Value, idiosyncratic volatility, total volatility, and Fama-French 3-factor alphas and betas, which are measured by regressing each stock's daily excess returns on daily factors. The key independent variable of interest is the dummy "WSB option flag", which equals one if the stock ticker and words like "option", "call", or "put" are co-mentioned in the same comment on the same day, or zero otherwise. Fama-MacBeth regression results and panel regression results are reported separately in Panel A and Panel B.

Panel A: Fama-MacBeth regressions

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| VARIABLES | change in log_user | log_user | change in trade_vol | trade_vol |
| WSB_options_flag | 0.006*** | 1.237*** | 12.983*** | 2.189*** |
| | (0.001) | (0.110) | (1.012) | (0.424) |
| Controls | YES | YES | YES | YES |
| Observations | 1,209,101 | 1,212,063 | 1,212,063 | 1,212,063 |
| Avg R-squared | 0.050 | 0.421 | 0.223 | 0.073 |
| Number of groups | 600 | 601 | 601 | 601 |

Panel B: Panel regressions

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| VARIABLES | change in log_user | log_user | change in trade_vol | trade_vol |
| WSB_options_flag | 0.005*** | 1.545*** | 10.222*** | 1.829*** |
| | (0.000) | (0.127) | (1.400) | (0.161) |
| Controls | YES | YES | YES | YES |
| Observations | 1,209,101 | 1,212,063 | 1,212,063 | 1,212,063 |
| R-squared | 0.003 | 0.392 | 0.156 | 0.004 |
| Firm Cluster | YES | YES | YES | YES |
| Time Cluster | YES | YES | YES | YES |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 11: The Impact of Retail Participation on Monthly Crash Risk
This table reports the result of a difference-in-difference analysis for the impact of retail participation on ex-ante monthly firm-level crash risk. The dependent variable is the estimated ex-ante monthly crash risk from the Ridge model. *Treatment* is a dummy variable that equals one if both firm ticker and option terms are mentioned in comments in Reddit Wallstreetbets in 2018. *Post* is a dummy variable that equals one if the year is 2018. In Column (1) to (4), Standard errors are clustered at both firm and month levels. In Column (2) and (3), I add a plethora of firm characteristics; in Column (4), I add predictor variables used in estimating ex-ante monthly crash risk. Column (5) adds firm and time fixed effects. Sample runs from January 2017 to December 2018. The base specification (without fixed effects) is:

$$Crash\,Risk_{i,t+1} = \alpha + \beta_0 Treated + \beta_1 Post + \beta_2 Treated \times Post + \gamma Controls_{i,t} + \epsilon_{i,t}$$

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Dep Var: ex-ante Monthly Crash Risk | | | | |
| | Clustered | | | | FE |
| 1.treatment | -0.032*** | 0.011*** | 0.011*** | -0.001 | |
| | (0.005) | (0.003) | (0.003) | (0.002) | |
| 1.post | -0.087*** | -0.085*** | -0.085*** | -0.088*** | |
| | (0.028) | (0.029) | (0.029) | (0.029) | |
| 1.treatment#1.post | 0.016*** | 0.015*** | 0.015*** | 0.010*** | 0.009*** |
| | (0.003) | (0.003) | (0.003) | (0.002) | (0.001) |
| Size | | -0.043*** | -0.043*** | -0.023*** | -0.028*** |
| | | (0.003) | (0.003) | (0.002) | (0.002) |
| B2M | | -0.001 | -0.002 | -0.008 | -0.003 |
| | | (0.006) | (0.006) | (0.005) | (0.004) |
| ROE | | | -0.000*** | -0.000*** | -0.000* |
| | | | (0.000) | (0.000) | (0.000) |
| ATG | | | -0.004 | -0.007 | -0.008** |
| | | | (0.010) | (0.010) | (0.004) |
| Exret3 | | | -0.016 | -0.019 | -0.027*** |
| | | | (0.019) | (0.018) | (0.003) |
| Predictors | NO | NO | NO | YES | YES |
| Observations | 39,482 | 39,482 | 39,482 | 39,482 | 39,411 |
| R-squared | 0.120 | 0.416 | 0.416 | 0.532 | 0.868 |
| Firm Cluster | YES | YES | YES | YES | NO |
| Time Cluster | YES | YES | YES | YES | NO |
| Firm FE | NO | NO | NO | NO | YES |
| Time FE | NO | NO | NO | NO | YES |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 12: The Impact of Retail Participation on Crash Risk: Big vs Small Firms
This table reports the result of a triple difference-in-difference analysis for the impact of retail participation on ex-ante monthly firm-level crash risk for big and small firm cohorts. The dependent variable is the estimated ex-ante monthly crash risk from the Ridge model. *Treatment* is a dummy variable that equals one if both firm ticker and option terms are mentioned in comments in Reddit Wallstreetbets in 2018. *Post* is a dummy variable that equals one if the year is 2018. *Big* is a dummy variable that equals one if the firm is larger than the medium size at the beginning of the sample, or zero otherwise. In Column (1) to (3), Standard errors are clustered at both firm and month levels. In Column (2) and (3), I add a plethora of firm characteristics. Column (4) adds firm and time fixed effects. Sample runs from January 2017 to December 2018.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Dep Var: ex-ante Monthly Crash Risk | | | |
| | Clustered | | | FE |
| 1.treatment | 0.006 | 0.011*** | 0.012*** | |
| | (0.005) | (0.004) | (0.004) | |
| 1.post | -0.096*** | -0.090** | -0.090** | |
| | (0.032) | (0.033) | (0.033) | |
| 1.treatment#1.post | 0.014*** | 0.019*** | 0.018*** | 0.015*** |
| | (0.004) | (0.004) | (0.005) | (0.002) |
| 1.big | -0.108*** | -0.013* | -0.014* | |
| | (0.007) | (0.007) | (0.007) | |
| 1.treatment#1.big | -0.023*** | 0.001 | 0.001 | |
| | (0.007) | (0.005) | (0.005) | |
| 1.post#1.big | 0.016 | 0.012 | 0.012 | 0.012*** |
| | (0.014) | (0.013) | (0.014) | (0.001) |
| 1.treatment#1.post#1.big | -0.006 | -0.010** | -0.010** | -0.006** |
| | (0.004) | (0.004) | (0.004) | (0.003) |
| Size | | -0.041*** | -0.041*** | -0.051*** |
| | | (0.003) | (0.003) | (0.003) |
| B2M | | -0.001 | -0.002 | 0.007 |
| | | (0.006) | (0.006) | (0.006) |
| ROE | | | -0.000*** | -0.000* |
| | | | (0.000) | (0.000) |
| ATG | | | -0.005 | -0.007* |
| | | | (0.010) | (0.004) |
| Exret3 | | | -0.017 | -0.026*** |
| | | | (0.019) | (0.003) |
| Observations | 39,482 | 39,482 | 39,482 | 39,411 |
| R-squared | 0.298 | 0.416 | 0.417 | 0.840 |
| Firm Cluster | YES | YES | YES | NO |
| Time Cluster | YES | YES | YES | NO |
| Firm FE | NO | NO | NO | YES |
| Time FE | NO | NO | NO | YES |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 13: The Impact of Retail Participation on Crash Risk: PSM Approach
This table reports the result of various difference-in-difference analyses for the impact of retail participation on ex-ante monthly firm-level crash risk for big and small firm cohorts, by using propensity score matching. The dependent variable is the estimated ex-ante monthly crash risk from the Ridge model. *Treatment* is a dummy variable that equals one if both firm ticker and option terms are mentioned in comments in Reddit Wallstreetbets in 2018. *Post* is a dummy variable that equals one if the year is 2018. *Big* is a dummy variable that equals one if the firm is larger than the medium size at the beginning of the sample, or zero otherwise. Each treatment stock is matched with at least one control firm, based on propensity score matching. The propensity scores are generated by logistic regression of treatment dummy on firm characteristics at the beginning of the sample. In Column (4) and (5), control variables are added. Sample runs from January 2017 to December 2018.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | Dep Var: ex-ante Monthly Crash Risk | | | | |
| | PSM matched | | | | |
| 1.treatment | -0.004 | 0.010 | | 0.014** | |
| | (0.005) | (0.007) | | (0.006) | |
| 1.post | -0.087*** | -0.101*** | | -0.095*** | |
| | (0.026) | (0.031) | | (0.031) | |
| 1.treatment#1.post | 0.017*** | 0.018*** | 0.010*** | 0.022*** | 0.015*** |
| | (0.003) | (0.006) | (0.003) | (0.006) | (0.003) |
| 1.big | | -0.109*** | | -0.018** | |
| | | (0.009) | | (0.008) | |
| 1.treatment#1.big | | -0.022** | | -0.008 | |
| | | (0.009) | | (0.006) | |
| 1.post#1.big | | 0.023* | 0.018*** | 0.019 | 0.015*** |
| | | (0.013) | (0.002) | (0.013) | (0.002) |
| 1.treatment#1.post#1.big | | -0.014** | -0.006 | -0.017** | -0.009*** |
| | | (0.006) | (0.003) | (0.006) | (0.003) |
| Controls | NO | NO | NO | YES | YES |
| Observations | 19,584 | 19,584 | 19,574 | 19,584 | 19,574 |
| R-squared | 0.111 | 0.322 | 0.832 | 0.449 | 0.844 |
| Firm Cluster | YES | YES | NO | YES | NO |
| Time Cluster | YES | YES | NO | YES | NO |
| Firm FE | NO | NO | YES | NO | YES |
| Time FE | NO | NO | YES | NO | YES |

*Note:* $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

Table 14: The Impact of Retail Participation on Crash Related Characteristics
This table reports the result of a triple difference-in-difference analysis for the impact of retail participation on ex-ante monthly characteristics for big and small firm cohorts. The dependent variables include: the relative deep out-of-money put and call option prices; trading volume as volume scaled by shares outstanding; total return volatility; and total return skewness. *Treatment* is a dummy variable that equals one if both firm ticker and option terms are mentioned in comments in Reddit Wallstreetbets in 2018. *Post* is a dummy variable that equals one if the year is 2018. *Big* is a dummy variable that equals one if the firm is larger than the medium size at the beginning of the sample, or zero otherwise. Standard errors are clustered at both firm and month levels. Sample runs from January 2017 to December 2018.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | | | Dep Vars: | | |
| VARIABLES | FOMP | FOMC | Trade_Vol | Tvol | Tskew |
| 1.treatment | 0.004*** | 0.005*** | 1.327*** | 0.003*** | 0.025 |
| | (0.001) | (0.001) | (0.256) | (0.001) | (0.017) |
| 1.post | 0.001*** | 0.000 | 0.041 | 0.003** | -0.037 |
| | (0.000) | (0.001) | (0.088) | (0.001) | (0.042) |
| 1.treatment#1.post | 0.003*** | 0.003*** | 0.332 | 0.003*** | 0.003 |
| | (0.001) | (0.001) | (0.238) | (0.001) | (0.027) |
| 1.big | 0.002** | 0.001 | 0.588*** | -0.000 | -0.011 |
| | (0.001) | (0.001) | (0.125) | (0.001) | (0.041) |
| 1.treatment#1.big | -0.000 | -0.001 | -0.865*** | -0.001 | -0.021 |
| | (0.001) | (0.002) | (0.274) | (0.001) | (0.035) |
| 1.post#1.big | -0.000 | -0.000 | 0.061 | 0.001 | -0.011 |
| | (0.000) | (0.000) | (0.058) | (0.001) | (0.058) |
| 1.treatment#1.post#1.big | -0.002** | -0.003** | -0.230 | -0.002** | -0.017 |
| | (0.001) | (0.001) | (0.226) | (0.001) | (0.041) |
| Controls | YES | YES | YES | YES | YES |
| Observations | 39,482 | 39,482 | 39,482 | 39,482 | 39,482 |
| R-squared | 0.350 | 0.339 | 0.062 | 0.213 | 0.078 |
| Firm Cluster | YES | YES | YES | YES | YES |
| Time Cluster | YES | YES | YES | YES | YES |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table 15: Decile High-Minus-Low Alphas

Stocks are ranked by their ex-ante crash probabilities each month into ten deciles. This table presents the high-minus-low long-short zero-cost strategy alphas, per model, for both equal-weighted and value-weighted portfolios. $t$-statistics are in the even rows. Time-series regressions are estimated with Newey-West standard errors with 12 lags.

| model | Logit | LASSO | Ridge | Elastic Net | Linear SVM | Logit | LASSO | Ridge | Elastic Net | Linear SVM |
|---|---|---|---|---|---|---|---|---|---|---|
| | Equal-Weighted | | | | | Value-Weighted | | | | |
| CAPM | -1.587 | -1.532 | -1.379 | -1.519 | -1.695 | -1.665 | -1.627 | -1.488 | -1.473 | -1.626 |
| t | -3.144 | -2.842 | -2.517 | -2.995 | -3.528 | -3.127 | -3.027 | -2.809 | -2.802 | -3.068 |
| FF3 | -1.526 | -1.476 | -1.311 | -1.466 | -1.636 | -1.611 | -1.578 | -1.434 | -1.433 | -1.565 |
| t | -4.737 | -3.678 | -3.328 | -3.921 | -5.109 | -4.421 | -3.757 | -3.481 | -3.471 | -4.208 |
| FF4 | -1.149 | -1.124 | -0.965 | -1.105 | -1.268 | -1.200 | -1.199 | -1.062 | -1.029 | -1.172 |
| t | -3.692 | -3.102 | -2.697 | -3.418 | -3.972 | -3.316 | -2.926 | -2.690 | -2.709 | -3.151 |
| FF5 | -0.770 | -1.013 | -0.823 | -0.989 | -0.936 | -0.770 | -0.994 | -0.837 | -0.792 | -0.764 |
| t | -2.718 | -2.615 | -2.289 | -2.591 | -3.505 | -2.452 | -2.710 | -2.383 | -2.012 | -2.392 |
| FF6 | -0.537 | -0.786 | -0.600 | -0.756 | -0.706 | -0.517 | -0.752 | -0.602 | -0.536 | -0.522 |
| t | -2.289 | -2.442 | -1.970 | -2.531 | -3.090 | -1.867 | -2.280 | -1.933 | -1.623 | -1.776 |

Table 16: Out-of-Money Option Trading Around Experiment
The table examines whether there is significant increase in out-of-money option trading volume after the introduction of commission-free option trading by Robinhood. The tests use the same difference-in-difference specifications employed in the retail trading section. Control groups are selected via PSM matching. Control variables include log market value of equity, log underlying stock trading volume, log price, rolling one-month stock return volatility, and one-day lagged stock return. All control variables are lagged one day. All variables are measured at daily frequency.

| Panel A: Out-of-Money Put Volume | | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| VARIABLES | log_volp | log_volp | log_volp | log_volp |
| 1.treatment | 0.507*** | | 0.251** | |
| | (0.079) | | (0.107) | |
| 1.post | -0.141*** | | -0.122** | |
| | (0.040) | | (0.052) | |
| 1.treatment#1.post | 0.112*** | 0.157*** | 0.145** | 0.245*** |
| | (0.039) | (0.009) | (0.066) | (0.015) |
| 1.big_all | | | -0.641*** | |
| | | | (0.127) | |
| 1.treatment#1.big_all | | | 0.349** | |
| | | | (0.151) | |
| 1.post#1.big_all | | | -0.034 | -0.008 |
| | | | (0.059) | (0.013) |
| 1.treatment#1.post#1.big_all | | | -0.090 | -0.151*** |
| | | | (0.080) | (0.019) |
| Controls | YES | YES | YES | YES |
| Observations | 429,320 | 429,320 | 429,320 | 429,320 |
| R-squared | 0.502 | 0.701 | 0.507 | 0.701 |
| Firm & Time Cluster | YES | NO | YES | NO |
| Firm & Time FE | NO | YES | NO | YES |
| Panel B: Out-of-Money Call Volume | | | | |
| | (1) | (2) | (3) | (4) |
| VARIABLES | log_volc | log_volc | log_volc | log_volc |
| 1.treatment | 0.497*** | | 0.287** | |
| | (0.087) | | (0.116) | |
| 1.post | 0.057 | | -0.030 | |
| | (0.042) | | (0.058) | |
| 1.treatment#1.post | 0.141*** | 0.215*** | 0.226*** | 0.341*** |
| | (0.046) | (0.009) | (0.077) | (0.015) |
| 1.big_all | | | -0.491*** | |
| | | | (0.140) | |
| 1.treatment#1.big_all | | | 0.298* | |
| | | | (0.165) | |
| 1.post#1.big_all | | | 0.136* | 0.158*** |
| | | | (0.069) | (0.013) |
| 1.treatment#1.post#1.big_all | | | -0.151 | -0.203*** |
| | | | (0.095) | (0.019) |
| Controls | YES | YES | YES | YES |
| Observations | 429,320 | 429,320 | 429,320 | 429,320 |
| R-squared | 0.515 | 0.720 | 0.517 | 0.720 |
| Firm & Time Cluster | YES | NO | YES | NO |
| Firm & Time FE | NO | YES | NO | YES |

*Note:* *p<0.1; **p<0.05; ***p<0.01

# Appendix A. Complex Models

This section provides results from using more complex machine learning models to predict monthly crash risk. Here complexity refers to both number of independent variables (features) and the underlying machine learning model. In terms of features, the set is enlarged to include 134 variables in total, the vast majority of which consist of option-related variables. Section 7 gives detailed construction of the features. In terms of underlying machine learning models, two more models are explored: XGBoost tree methods and feed forward neural network.

Before feeding data into machine learning models, the following pre-processing for each window is conducted: first, standardize the training data, and use the information to transform the validation and test data into standardized data; second, extract principal components (either 5, 10, or 20) from training data, then use the information to extract the same number of PCs from validation and test data; finally, apply SMOTE to balance the training data. Then the training data is fed into each of the machine learning models to train the model; then I tune hyper-parameters using the validation data to find the best estimator, and finally fit the test data using the best estimator. The tuning parameters for each model are shown in Table A.1.

[Table A.1 about here.]

I report the mean performance metrics for all three machine learning models in the complex setting using either 5, 10, or 20 principal components, as well as the simple ridge regression used as the main results in Table A.2.

[Table A.2 about here.]

As shown in Table A.2, compared with main results in Column (3), the results from more complex models are quantitatively similar. To main parsimony and interpretability, I show the simple model as the main results in Section 4.

# Appendix B. Replications

I replicate the main results of Jang and Kang (2019) for the sample period 1996 - 2019. First, I confirm the main results of multinomial logistic regression of exploring the relationship between crashes and jackpots and various firm characteristics. Table A.3 show the results that are fairly consistent with the original test.

[Table A.3 about here.]

I then use the expanding training data to run the multinomial regressions and then predict one-year-ahead probability of crashes and jackpots out-of-sample. Starting from 4 years of training sample, the prediction window starts from January 2001 and ends at December 2019. For each month, I form high-minus-low portfolios by sorting stocks based on the predicted crash probabilities into deciles, and then regress either equally weighted or value weighted portfolio returns on CAPM, Fama-French 3-, 4-, 5- and 6-factor models. Table A.4 show the resulting alphas and associated t-statistics estimated using Newey-West standard errors with 12 lags (Newey and West (1986)).

[Table A.4 about here.]

As the table shows, the results from value-weighted portfolios on CAPM, and FF3 and FF4 models are consistent with Jang and Kang (2019). However, they are no longer significant when FF5 and FF6 factors are used, and they are not significant under the equally weighted scheme.

# Appendix C. Selected Variable Definitions

$ACI$ = CAPX ratio increase over the previous three periods mean. CAPX ratio is $CAPX/SALE$.

$AG$ = asset growth over the previous year

$Book\_value\_equity$ = $SEQ + TXDITC - Perferred$, preferred is $PSTKRV$, or $PSTKL$, or $PSTK$, whichever is first available.

$Crash\_Risk$ = predicted monthly ex ante probability of stock crash, with log return less than -20%

$FOMC$ = ratio between deep out-of-money call option price and the underlying implied forward stock price

$FOMP$ = ratio between deep out-of-money put option price and the underlying implied forward stock price

$GP$ = gross profitability, equals $(REVT - COGS)/AT$

$Illiquidity$ = monthly mean of daily absolute return over price times volume of that day, see Amihud (2002).

$Jackpot\_Risk$ = predicted monthly ex ante probability of jackpot, with log return greater than 20%

$NOA$ = $net\_operating\_assets/lag\_AT$

$NSI$ = natural log of changes in adjusted shares

$OSCR$ = $-1.32 - 0.407 \times ASIZE + 6.03 \times TLTA - 1.43 \times WCTA + 0.0757 \times CLCA - 1.72 \times OENEG - 2.37 \times NITA - 1.83 \times FUTL + 0.285 \times INTWO - 0.521 \times CHIN$, O-score, see Ohlson (1980).

$ROA$ = $NI/AT$

$SMIRK$ = difference between the implied volatility of out-of-money put option and at-the-money call option, see Xing et al. (2010)

$Tang$ = $PPENT/AT$

Table A.1: Tuning Parameters
The table shows the hyper-parameters for each underlying machine learning model in the complex setting, where there are 134 features (independent variables). The machine learning models include ridge regression, XGBoost tree methods, and feed forward neural network (NN, or multi-layer perceptrons, or MLP).

| | Parameter | Range |
|---|---|---|
| Ridge | $\lambda$ | 100 $\lambda$s between $10^{-4}$ and $10^4$ |
| XGBoost | max depth | 1,2,3,4,5,6 |
| | learning rate | 0.001 |
| | early stopping | 10 rounds |
| MLP (NN) | $\lambda$ | 100 $\lambda$s between $10^{-4}$ and $10^4$ |
| | hidden layers | (50,), (50,50), or (50,50,50) |

Table A.2: Mean Performance Metrics for Simple and Complex Models

The table reports mean performance metrics for the simple ridge regression used as main results, and three machine learning models in complex setting using either 5, 10, or 20 principal components, across the rolling prediction windows from January 1996 to December 2019. In all cases, each window consists of 5 months of training data, 1 month of validation data, and 1 month of test data, where hyper-parameters are tuned through validation data, and then the best estimator is chosen to fit the test set. The metrics are defined as follows:

$$Precision = \frac{True\,Positives}{True\,Positives + False\,Positives}$$
$$Recall = \frac{True\,Positives}{True\,Positives + False\,Negatives}$$
$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

These metrics are computed for each of the three classes. There are in total 281 windows, and hence 281 sets of metrics are generated in total for each underlying model. These metrics are then averaged across time.

| | | Main | Complex Models | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Simple | 5 PCs | | | 10 PCs | | | 20 PCs | | |
| Class | Metrics | Ridge | Ridge | XGBoost | NN | Ridge | XGBoost | NN | Ridge | XGBoost | NN |
| Crash | precision | 0.128 | 0.118 | 0.108 | 0.099 | 0.115 | 0.106 | 0.099 | 0.114 | 0.101 | 0.094 |
| | recall | 0.412 | 0.328 | 0.339 | 0.251 | 0.328 | 0.354 | 0.217 | 0.328 | 0.364 | 0.171 |
| | F1 | 0.128 | 0.138 | 0.130 | 0.113 | 0.136 | 0.133 | 0.107 | 0.136 | 0.132 | 0.098 |
| Plain | precision | 0.935 | 0.926 | 0.933 | 0.917 | 0.927 | 0.931 | 0.910 | 0.926 | 0.927 | 0.903 |
| | recall | 0.626 | 0.729 | 0.603 | 0.705 | 0.716 | 0.597 | 0.744 | 0.708 | 0.593 | 0.785 |
| | F1 | 0.730 | 0.812 | 0.724 | 0.790 | 0.804 | 0.721 | 0.811 | 0.799 | 0.717 | 0.834 |
| Jackpot | precision | 0.090 | 0.087 | 0.073 | 0.074 | 0.081 | 0.073 | 0.071 | 0.084 | 0.076 | 0.072 |
| | recall | 0.344 | 0.304 | 0.376 | 0.234 | 0.300 | 0.374 | 0.187 | 0.303 | 0.348 | 0.153 |
| | F1 | 0.108 | 0.112 | 0.105 | 0.092 | 0.109 | 0.104 | 0.083 | 0.111 | 0.105 | 0.080 |

Table A.3: Replication of Jang and Kang (2019)
The table replicates the multinomial logit regression from Jang and Kang (2019) for sample period 1996 - 2019. Variable definitions follow the paper referenced. Standard errors are clustered at stock and month levels and are included in parentheses.

| | Crash | | Jackpot | |
|---|---|---|---|---|
| | Coefficient | | Coefficent | |
| rm12 | 1.038*** | (0.294) | -0.994*** | (0.244) |
| exret12 | -0.191*** | (0.0509) | -0.211*** | (0.0398) |
| tvol | 28.53*** | (1.761) | 25.20*** | (1.178) |
| tskew | 0.0323*** | (0.00680) | 0.0217*** | (0.00766) |
| size | -0.00731 | (0.0121) | -0.154*** | (0.0156) |
| dturn | -0.0238 | (0.0494) | -0.315*** | (0.0554) |
| age | -0.0222*** | (0.00177) | -0.0149*** | (0.00171) |
| tang | 0.121*** | (0.0352) | 0.119*** | (0.0320) |
| salesg | 0.200*** | (0.0237) | 0.0375 | (0.0263) |
| Constant | -2.968*** | (0.273) | -0.511 | (0.325) |
| Observations | 965,401 | | Pseudo R2 | 0.102 |

*Note:* *p<0.1; **p<0.05; ***p<0.01

Table A.4: High-Minus-Low Alphas per Jang and Kang (2019)
The table presents the results from regressing high-minus-low crash risk portfolio returns on various asset pricing factors, following Jang and Kang (2019). Each month, I sort stocks into deciles based on the predicted crash probabilities, and then calculate either equally weighted or value weighted portfolio returns. Then the time series of returns are regressed on time series of asset pricing factors. The sample runs from January 2001 to December 2019. Standard errors are Newey-West standard errors with 12 lags.

| | Equal-weighted | | Value-weighted | |
|---|---|---|---|---|
| pricing model | alpha | T-stat | alpha | T-stat |
| CAPM | -0.348 | -0.624 | -1.078*** | -2.814 |
| FF3 | -0.400 | -0.920 | -1.067*** | -3.487 |
| FF4 | -0.128 | -0.335 | -0.866** | -2.635 |
| FF5 | 0.583 | 1.156 | 0.088 | 0.256 |
| FF6 | 0.571 | 1.458 | 0.081 | 0.259 |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

# References


Abreu, D., Brunnermeier, M. K., 2003. Bubbles and crashes. Econometrica 71, 173–204.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of financial markets 5, 31–56.

An, H., Zhang, T., 2013. Stock price synchronicity, crash risk, and institutional investors. Journal of Corporate Finance 21, 1–15.

Andreou, P. C., Antoniou, C., Horton, J., Louca, C., 2016. Corporate governance and firm-specific stock price crashes. European Financial Management 22, 916–956.

Anthony, J. H., 1988. The interrelation of stock and options market trading-volume data. The Journal of Finance 43, 949–964.

Bali, T. G., Cakici, N., Whitelaw, R. F., 2011. Maxing out: Stocks as lotteries and the cross-section of expected returns. Journal of financial economics 99, 427–446.

Banerji, G., 2021. Individuals embrace options trading, turbocharging stock markets. Wall Street Journal, URL: https://www.wsj.com/articles/individuals-embrace-options-trading-turbocharging-stock-markets-11632661201?mod=hp_lead_pos4.

Barber, B. M., Huang, X., Odean, T., Schwarz, C., 2020. Attention induced trading and returns: Evidence from robinhood users. Available at SSRN 3715077 .

Barber, B. M., Odean, T., 2000. Trading is hazardous to your wealth: The common stock investment performance of individual investors. The journal of Finance 55, 773–806.

Barro, R. J., Liao, G. Y., 2020. Rare disaster probability and options pricing. Journal of Financial Economics .

Bates, D. S., 1991. The crash of '87: was it expected? the evidence from options markets. The journal of finance 46, 1009–1044.

Batista, G. E., Prati, R. C., Monard, M. C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter 6, 20–29.

Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond risk premiums with machine learning. The Review of Financial Studies 34, 1046–1089.

Callen, J. L., Fang, X., 2015. Short interest and stock price crash risk. Journal of Banking & Finance 60, 181–194.

Campbell, J. Y., Hilscher, J., Szilagyi, J., 2008. In search of distress risk. The Journal of Finance 63, 2899–2939.

Carhart, M. M., 1997. On persistence in mutual fund performance. The Journal of finance 52, 57–82.

Chang, X. S., Chen, Y., Zolotoy, L., 2016. Stock liquidity and stock price crash risk. Journal of Financial and Quantitative Analysis (JFQA), Forthcoming .

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Chen, J., Hong, H., Stein, J. C., 2001. Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. Journal of financial Economics 61, 345–381.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chu, Y., Hirshleifer, D., Ma, L., 2020. The causal effect of limits to arbitrage on asset pricing anomalies. The Journal of Finance 75, 2631–2672.

Conrad, J., Kapadia, N., Xing, Y., 2014. Death and jackpot: Why do individual investors hold overpriced stocks? Journal of Financial Economics 113, 455–475.

Cooper, M. J., Gulen, H., Schill, M. J., 2008. Asset growth and the cross-section of stock returns. the Journal of Finance 63, 1609–1651.

Daniel, K., Titman, S., 2006. Market reactions to tangible and intangible information. The Journal of Finance 61, 1605–1643.

De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J., 1990a. Noise trader risk in financial markets. Journal of political Economy 98, 703–738.

De Long, J. B., Shleifer, A., Summers, L. H., Waldmann, R. J., 1990b. Positive feedback investment strategies and destabilizing rational speculation. the Journal of Finance 45, 379–395.

Dechow, P. M., Sloan, R. G., Sweeney, A. P., 1995. Detecting earnings management. Accounting review pp. 193–225.

Diether, K. B., Lee, K.-H., Werner, I. M., 2009. It's sho time! short-sale price tests and market quality. The Journal of Finance 64, 37–73.

Elliott, G., Timmermann, A., 2016. Forecasting in economics and finance. Annual Review of Economics 8, 81–110.

Erel, I., Stern, L. H., Tan, C., Weisbach, M. S., 2021. Selecting directors using machine learning. The Review of Financial Studies 34, 3226–3264.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of .

Fama, E. F., French, K. R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fama, E. F., French, K. R., 2020. Comparing cross-section and time-series factor models. The Review of Financial Studies 33, 1891–1926.

Fama, E. F., MacBeth, J. D., 1973. Risk, return, and equilibrium: Empirical tests. Journal of political economy 81, 607–636.

Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: A test of new factors. The Journal of Finance 75, 1327–1370.

Feuer, W., 2021. First-time investors now make up 15% of retail market. Institutional Investor, URL: https://www.institutionalinvestor.com/article/b1r9ycrxwlld7j/First-Time-Investors-Now-Make-Up-15-of-Retail-Market.

Foucault, T., Sraer, D., Thesmar, D. J., 2011. Individual investors and volatility. The Journal of Finance 66, 1369–1406.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning, vol. 1. Springer series in statistics New York.

Graham, J. R., Kumar, A., 2006. Do dividend clienteles exist? evidence on dividend preferences of retail investors. The Journal of Finance 61, 1305–1336.

Greenwood, R., Nagel, S., 2009. Inexperienced investors and bubbles. Journal of Financial Economics 93, 239–258.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Han, B., Kumar, A., 2013. Speculative retail trading and asset prices. Journal of Financial and Quantitative Analysis 48, 377–404.

Herskovic, B., Kelly, B., Lustig, H., Van Nieuwerburgh, S., 2016. The common factor in idiosyncratic volatility: Quantitative asset pricing implications. Journal of Financial Economics 119, 249–283.

Hirshleifer, D., Hou, K., Teoh, S. H., Zhang, Y., 2004. Do investors overvalue firms with bloated balance sheets? Journal of Accounting and Economics 38, 297–331.

Hutton, A. P., Marcus, A. J., Tehranian, H., 2009. Opaque financial reports, r2, and crash risk. Journal of financial Economics 94, 67–86.

Jang, J., Kang, J., 2019. Probability of price crashes, rational speculative bubbles, and the cross-section of stock returns. Journal of Financial Economics 132, 222–247.

Jin, L., Myers, S. C., 2006. R2 around the world: New theory and new tests. Journal of financial Economics 79, 257–292.

Kelley, E. K., Tetlock, P. C., 2017. Retail short selling and stock prices. The Review of Financial Studies 30, 801–834.

Kelly, B., Jiang, H., 2014. Tail risk and asset prices. The Review of Financial Studies 27, 2841–2871.

Kim, J.-B., Li, Y., Zhang, L., 2011. Corporate tax avoidance and stock price crash risk: Firm-level analysis. Journal of Financial Economics 100, 639–662.

Kim, Y., Li, H., Li, S., 2014. Corporate social responsibility and stock price crash risk. Journal of Banking & Finance 43, 1–13.

King, G., Zeng, L., 2001. Logistic regression in rare events data. Political analysis 9, 137–163.

Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. Journal of Financial Economics 135, 271–292.

Kumar, A., 2009. Who gambles in the stock market? The Journal of Finance 64, 1889–1933.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics 45, 221–247.

McCabe, C., 2020. New army of individual investors flexes its muscle. Wall Street Journal, URL: https://www.wsj.com/articles/new-army-of-individual-investors-flexes-its-muscle-11609329600.

McCrank, J., 2021. Factbox: The u.s. retail trading frenzy in numbers. Thomson Reuters, URL: https://www.reuters.com/article/us-retail-trading-numbers-idUSKBN29Y2PW.

Merton, R. C., 1973. An intertemporal capital asset pricing model. Econometrica: Journal of the Econometric Society pp. 867–887.

Newey, W. K., West, K. D., 1986. A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Novy-Marx, R., 2013. The other side of value: The gross profitability premium. Journal of Financial Economics 108, 1–28.

Ohlson, J. A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18, 109–131.

Pan, J., 2002. The jump-risk premia implicit in options: Evidence from an integrated time-series study. Journal of financial economics 63, 3–50.

Petersen, M. A., 2009. Estimating standard errors in finance panel data sets: Comparing approaches. The Review of Financial Studies 22, 435–480.

Pontiff, J., 1996. Costly arbitrage: Evidence from closed-end funds. The Quarterly Journal of Economics 111, 1135–1151.

Ripley, B. D., 1996. Pattern recognition and neural networks. Cambridge university press.

Ritter, J. R., 1991. The long-run performance of initial public offerings. The journal of finance 46, 3–27.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., Dormann, C. F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.

Robinhood, 2017. Introducing options trading. Robinhood Financial LLC, URL: `https://blog.robinhood.com/news/2017/12/12/introducing-options-trading`.

Seliya, N., Khoshgoftaar, T. M., Van Hulse, J., 2009. A study on the relationships of classifier performance metrics. In: *2009 21st IEEE international conference on tools with artificial intelligence*, IEEE, pp. 59–66.

Shleifer, A., Vishny, R. W., 1997. The limits of arbitrage. The Journal of finance 52, 35–55.

Titman, S., Wei, K. C. J., Xie, F., 2004. Capital Investments and Stock Returns. Journal of Financial and Quantitative Analysis 39, 677–700.

Welch, I., 2020. Retail raw: Wisdom of the robinhood crowd and the covid crisis. Tech. rep., National Bureau of Economic Research.

Xing, Y., Zhang, X., Zhao, R., 2010. What does the individual option volatility smirk tell us about future equity returns? Journal of Financial and Quantitative Analysis pp. 641–662.

Yan, S., 2011. Jump risk, stock returns, and slope of implied volatility smile. Journal of Financial Economics 99, 216–233.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67, 301–320.

Zweig, J., 2020. When the stock market is too much fun. Wall Street Journal, URL: `https://www.wsj.com/articles/when-the-stock-market-is-too-much-fun-11607705516`.