MIE368 Analytics in Action

Final Report

IPO Success Forecast & Investment Optimization

Yimamu(Elzat) Yilizhati

1. Introduction

The Initial Public Offering (IPO) is the process when a private company sells shares to the public for the first time. Often, the company's expectation can be significantly different from the market reaction which leads to a huge loss of capital. Hence, private companies need some metrics to decide if it is the right time to go public, ideally, a strong predictive model that can forecast the rate of return of the IPO. Meanwhile, the investors need a model that selects the best investment portfolio according to different objectives: risk-seeking, risk-averse or balanced.

The main goal of this project consists of two steps: 1. Predicting the rate of return of each IPO on the releasing date using random forest regressor 2. Optimizing a portfolio that constructs a distribution of investment in private companies based on the type of investors.

Our team collected the company's features and market variables that traditionally seen as good predictors to the closing price of IPO. Then we took in those features to build a statistical model to predict the closing price on the day of the IPO releases. The rate of return of each private company can be calculated with the predicted value of $\frac{(close\ price - offer\ price)}{offer\ price}$ and the risk is quantified using the standard deviation. Then, a set of companies' rate of return and risk are inputted into our portfolio optimization models that constructs an optimal allocation of investment on the companies based on investors' risk tolerance level.

2. Data

The datasets are sourced from Yahoo Finance[1] and IPOscoop[2] websites which contain approximately 500 records of basic company information including the company name, IPO offer price, estimated volume of shares, IPO date, first-day close price. Beyond these features, the team manually collected other 17 internal and external features including company's employee size, company age, industry, country, exchange platform (NASDAQ or NYSE), financial state(e.g. revenues & net income in million US dollars), estimated volume (in millions of shares), and whether it is served by the top 8 investment bank[1], as well as the economic trend of the market (i.e. S&P 500 index, gold price, previous industry IPO return and US prime rate).

2.1 EDA - Correlation Analysis

The correlation between the target variable (first-day close price) and each feature is evaluated through correlation analysis. (Full table in Appendix A and B). As a result, the team classified the top 3 features that have the highest correlation with the target variable as "strong predictor" and 3 features with the lowest correlation as "weak predictor". To enhance the weak predictors, feature engineering is implemented to check if any stronger features can be created based on those weak features. Otherwise, these weak features are dropped to improve the models.

Strong Predictors	Weak Predictors
Offer price	Age of the company

Whether the company is served by a top 8 investment bank	Net income of the company
Whether the company is US-based	Employee size of the company

Table 1 - Top strong and weak predictors

Multicollinearity can hinder the interpretability and performance of our models as the features are not independent. Additionally, collinearity violates the assumption of Linear Regression which is one of our potential prediction models. To address this problem, the correlations among all the features are visualized in order to find highly correlated features (See Appendix C). The pairs of highly correlated features are 'the number of shares' & 'estimated volume of the company', 'company revenue' & 'estimated volume of the company' and 'employee size' & 'company revenue'. The team decided to further analyze those features and see which features we need to drop.

2.2 EDA - Distribution Analysis

Based on the correlation analysis above, all the features that have strong correlations with the target variable (first-day close price) are plotted in the empirical distribution for checking normality.

For instance, the distribution plot for 'Offer Price' which has the highest correlation value (0.817) with the target variable, Figure 1, shows an approximately normal distribution. And the majority offer price values are in the range of (0,50) with a mean of $17.5 and a standard deviation of $7.25. All the other variables which have correlations with the target variable are plotted in Appendix D. The target variable, 'first-day close price' is plotted in Figure 2 which also shows approximately normal distribution with right skewness and mean of $18.7.

Based on the findings of variables' normality, the data transformation such as log transformation and square root transformation will be applied to solve the extreme skewness.

The dataset collected has no missing data. Additionally, no significant outlier has been spotted according to outlier analysis (Appendix E).

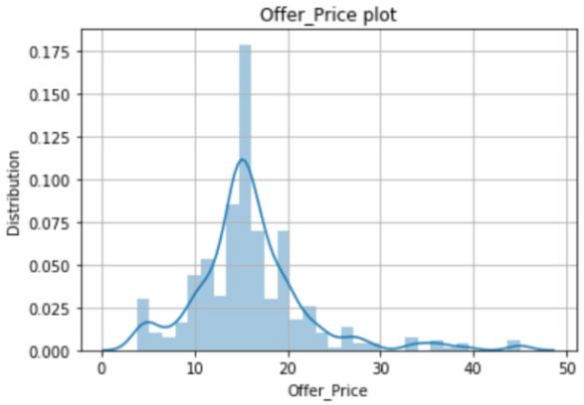


Figure 1-Distribution plot of Offer Price

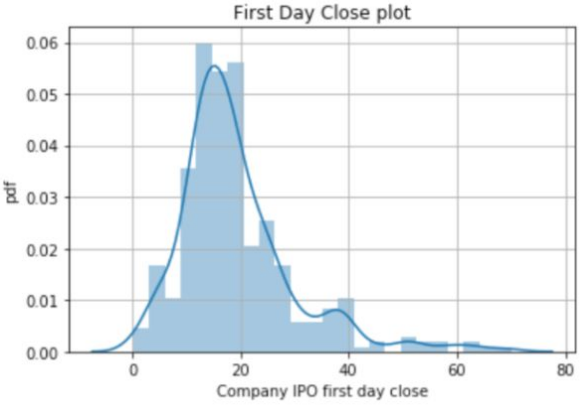


Figure 2-Distribution plot of First Day Close Price

3.0 Method

In this section, the team first described the processes of how the team discovered the best prediction model for the first-day closing price as well as the professional techniques that the team used to boost the performance of the models. The team then discussed the strategies of how to analyze the elements in the best prediction model to generate the risk and rate of return on each company's IPO. With the risk and rate of return as input information into optimization modelling, the team illustrated how the portfolio optimization model was built and how the model can help three different types of the investor on decision making.

3.1 Prediction Model

As the problem statement mentioned above, the team aimed to predict the first-day closing price of companies' IPO. Based on research on the IPO background and benchmarking among existing prediction models, the team chose linear regression (LR) and random forest regression (RFR) to train the dataset since LR is the traditional algorithm being used in the financial sector to discover the relationship between features and targeted value [4]. RFR was selected because the complexity of the IPO problem, and RFR, the nonlinear method, has proved to be more effective and accurate in the field of finance[4]. Moreover, the random forest regression can be used to determine the volatility of each company's first closing price.

At the beginning of the process, the team decided to split the former 80% data points as the training set and the latest 20% as the testing set according to the offering date, considering the temporal nature of the problem. The team used the raw dataset to train both LR and RFR models to set the baseline for each model as well as compare their accuracy results before other feature engineering techniques being applied. Based on results, the team then performed data engineering and feature engineering to both models with respect to their characteristics.

To enhance performance on both models, all continuous variables are scaled using standard scaling function as some features have vastly different orders of magnitude. Next, the team explored more features by experimenting with the interaction methods. In this case, the team tried multiplication, division, square root, and other possible operations, because of the lack of expertise in the financing field. The results of the interaction method with meaningful improvement are shown in Table 2. For example, by combining two weak features, employee size (ES) with 0.148 correlation and previous industry IPO return (PIR) with 0.071 correlation, the new feature ES*RIP is generated by the interaction method with 0.324 correlation which can be considered as a strong predictor in this case.

Feature 1	Feature 1 Correlation	Feature 2	Feature 2 Correlation	New Feature	New Correlation
Employee Size (ES)	0.148	Previous Industry IPO Return (PIR)	0.071	ES*PIR	0.324
Net Income in millions (NIM)	0.042	Shares in millions (SM)	0.089	NIM*SM	0.215

Gold price (GP)	0.152	Estimated volume in millions (EVM)	0.226	Sqrt(GP/EVM)	0.464

Table 2 - The new feature generated and the improvement results

According to the distribution analysis above, the team realized the extreme skewness of some features is against one of the linear regression's key assumptions, which is the normality of the data. Therefore, the team performed data transformation by normalizing the data by applying the log operation to spread the low-value data in a larger ranger. Table 3 below compares the skewness before and after applying log on the features.

Variable	Original absolute skewness	Absolute skewness after log
Estimated volume in millions	7.310	0.394
Revenues in millions	5.044	1.166

Table 3 - The skewness value before and after transformation for two example variables

When it comes to the improvement of the RFR model, the team used grid search methods to explore the best hyperparameters (e.g. max_features=0.5, max_depth=4, n_estimators=100) in the RFR model function based on their performance.

With all the potential improvement implemented, the team used the improved dataset to train both LR and RFR models with optimal parameters. Finally, as we discovered in the previous section, some features have very weak correlations to the target value. Hence, by comparing the performance before and after dropping a weak feature, some features that decrease the performance of the models are dropped to avoid overfitting issues.

To further explore the benefits of the RFR model, the team used a 'for loop' to collect the decisions of each tree (in total 100 trees which is the 'n_estimator' parameter in the RFR) made when they are all combined to predict one company's first-day closing price. The predicted first-day closing price would be the mean value of the 100 data points, which would be used to calculate the rate of return on each company in the optimization model. Also, the standard deviation of 100 data points would be considered as the risk of each company's IPO with respect to their rate of return.

3.3 Optimization Model

In order to construct an applicable solution that can assist investors, the team decided to build a portfolio optimization model. The linear programming (LP) model will take the predicted returns and the standard deviations from decision trees in the RFR model as inputs. Then, it will produce the optimal weighted asset proportion strategy based on constraints and objectives. The model should be able to satisfy different kinds of investors' needs according to their risk aptitudes.

According to the modern portfolio theory (MPT), the increase in potential asset return follows by the rise in return volatility. After analyzing the risk and return distribution produced by the prediction model, the team found the same MPT concept applies. In the model, the higher the predicted return means the greater the standard deviation among the decision trees. Hence, the trade-off between risk and return becomes a crucial decision to make as an investor. Since different investors have different expectations and risk tolerance levels, there is simply no one perfect investment strategy that works for every investor.

Generally, investors can be categorized as risk-averse, balanced and risk-seeking investors based on their risk tolerances. In our optimization model, each kind of investor has been associated with a risk threshold. The threshold level will serve as one of the constraints in the LP. Additionally, a constraint of maximum single asset weight is included for portfolio diversification purposes. Above constraint values are selected by the team to demonstrate the model performance in different scenarios, the numbers might not reflect reality. Based on the MPT, the common objective for all investors is maximizing their profit. Table 4 shows the risk tolerances and maximum asset weight the team defined for all investors. The flowchart in Figure 3 illustrates the workflow of the entire model.

Investor type	Risk tolerance	Maximum single asset weight
Risk-averse	15%	20%
Balanced	25%	30%
Risk-seeking	35%	50%

Table 4: Investor categorization

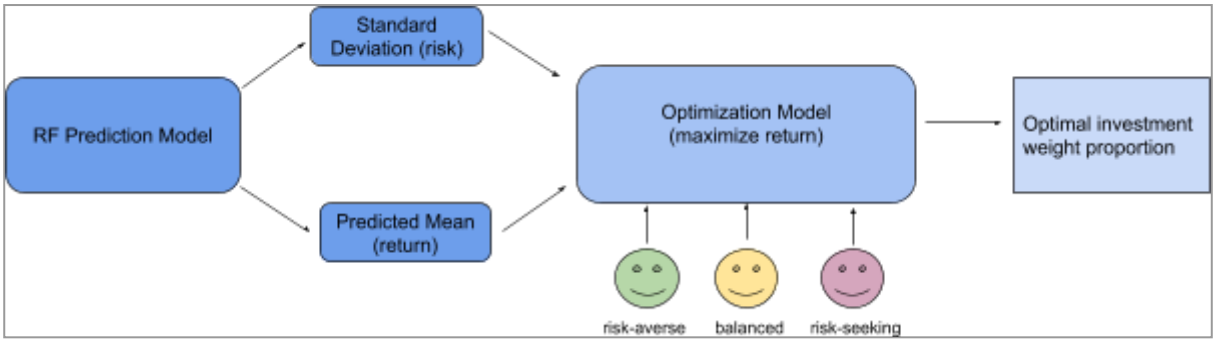


Figure 3: Optimization model flowchart

4. Results

The results could be divided into two sections, prediction results and optimization results. The prediction results will focus on the RFR model prediction accuracy and insights we discovered. In the optimization section, we will discuss how our model perform for different investor types and who benefits the most from our solution.

4.1 Prediction Result

For the baseline model trained by the raw dataset, the training and testing scores of LR and RFR models are recorded. Also, after adding the engineered features to the dataset, some highly correlated features are dropped to avoid multicollinearity, especially for the linear regression. In addition, all skewed features are log-transformed to the normalized data. The grid search has been performed to achieve the optimal accuracy of the models. Table 5 below shows the comparison of the baseline and improved models. The score represents the coefficient of determination which is calculated as $R^2 = \frac{1 - (sum\ of\ squared\ error)}{(sum\ of\ squared\ total)}$. The score value represents the proportion of explained variation over total variation. Hence, the score could reflect on how accurate the prediction model is.

Model Type	State	Train Score	Test Score
Linear Regression (LR)	Baseline	0.655	0.565
	Improved	0.654	0.565
Random Forest Regression (RFR)	Baseline	0.570	0.605
	Improved	0.792	0.733

Table 5 - Baseline and Improved Modelling Results

According to Table 5, the improved RFR model performed the best among all the model types and states, with the train score 0.792 and the test score 0.733. The 0.06 difference between train and test scores would be considered that the model is slightly overfitting. However, the overfitting issue could be resolved or minimized when taking more data points into the dataset.

4.2 Optimization Result

To demonstrate the optimization performance for three distinct investor types, the team first used the model to determine the optimal asset allocation strategies. For eliminating randomness in the result, we ran 100 replications of random company samples with size 30 for each investor type. The actual portfolio return for each investor type is calculated if they would follow the allocation that the model suggested, by using the equation $actual\ return = \sum_{i=1}^{30} optimized\ weight(i) * actual\ rate\ of\ return(i)$. Then, the average results of all replications are calculated and used to assess the model. The 95% confidence intervals are calculated to show the variation of the predicted returns as well. For the baseline, randomized weight allocation for each investor is generated, and the random return is calculated as $\sum_{i=1}^{30} random\ weight(i) * actual\ rate\ of\ return(i)$. Finally, we benchmarked the actual portfolio return of optimal strategy against the randomized portfolio return to show the improvement.

As shown in the graph below, Figure 4, the optimization model outperforms the baseline by around 3% in cases of risk-averse and balanced investor. A lesser performance, however, is shown for the risk-seeking investment strategy. The optimized risky portfolio produces a total return of 13.6% while the randomized investment allocation outputs 14.4% return.

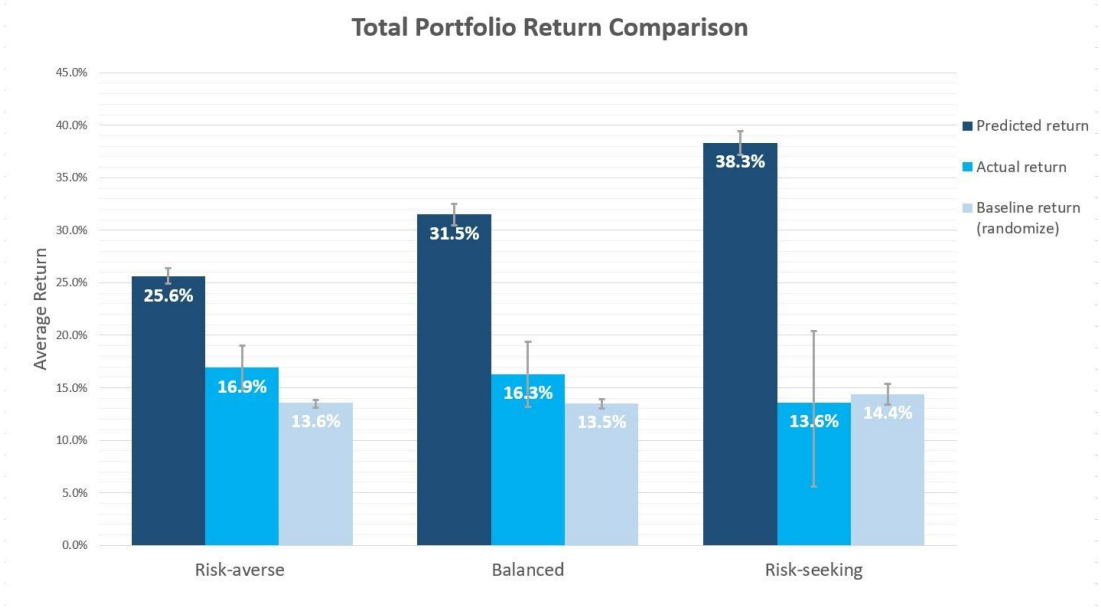


Figure 4: Total portfolio return comparison

Besides, the team discovered the overpredict tendency for all types of investors of our model. The predicted return is calculated as $\sum_{i=1}^{30} optimized\ weight(i) * predict\ rate\ of\ return(i)$. In the risk-seeking investor's case, the predicted return of 38.3% exceeds the actual return by around 25%. The 95% confidence level error bars also denotes the fact that the prediction accuracy diminishes rapidly as the risk level increases.

5. Discussion

In the result section above, we discovered many interesting insights regarding the model. In the prediction results, the team saw a much better performance from the random forest model compared to the linear regression model. This difference in prediction accuracy indicates that the IPO first-day close prices does not have a strong linear correlation with the selected features. This result does not serve as a surprise to the team as the IPO first-day close price heterogeneous nature is prevalent. Although random forest performs substantially better than the linear regression model, the team found that the model struggles to exceed the 0.733 testing score after multiple performance-boosting attempts. We deemed this issue is due to the lack of strong features for prediction. As we showed in the EDA section, the offer price has a significantly higher correlation with the target variable than the other predictors. The correlation of other variables did improve after the feature engineering step, however, offer price still acts as a dominant predictor.

According to the optimization result, the model shows a promising potential as improvements are obvious for risk-averse and balanced investors. Yet, our model is shown to be below the baseline performance once the portfolio type is switched to risk-seeking. We consider this fallout is largely due to the high single asset weight in the portfolio. As the risk tolerance is raised, the model tends to invest more on the asset with high-risk nature because of its promising return. Moreover, the asset weight on that risky IPO will also be exceedingly high as the maximum single asset is defined as 50%

for the risky investor. There is a high chance that one specific IPO company is extremely overpredicted by our model, this eventually leads to a poor total return on investment. The 95% confidence level error also proved our speculation, the error range dramatically increases in the high-risk category. Therefore, we highly recommend the built model only used by high risk-aversion investors with diversified portfolios.

6.0 Conclusion

The IPO provides speculative investors with a unique opportunity to grow their fortune by investing early in companies with great potential. Our goal is to help those investors to earn high returns from the initial bounce off the first day close price.

To predict the first-day close price, the team built a random forest regression model which shows a result of 73.3% coefficient of determination (R^2) on the testing set. In our next step of optimizing the asset allocation, our portfolio optimization model has outperformed the baseline (randomized) model by around 3% for risk-averse and balanced investors. However, risk-seeking investment has shown a less actual return compared to the baseline model. Based on those findings, we concluded that risk-averse investors who have lower risk tolerance level and tend to diversify their investment has a better chance of earning a profit off the bounce of the first day close price.

An interesting finding is that our prediction model forecasted that Uber will experience a 6.7% decrease in first-day close price which is close to the company's price drop (-8%) in reality on the first day of IPO release.

In conclusion, our prediction model and portfolio optimization model have proven to be able tackle the proposed problem and deliver a promising result. However, we have still spotted some room for future improvements which will be discussed in the next section.

7.0 Future Direction and Improvement

Although we have proved that the built solution can improve the asset allocation strategy for three types of investors, we still need to recognize the improvement from the baseline strategy is limited. To further enhance our solution, we should work on both aspects of the solution, prediction and optimization.

In terms of prediction, we can include more IPO company data records and strong features such as management expertise. This practice is to corporate more powerful predictors into our model. We deem this type of internal data can distinctively affect the public investors' confidence in the company's future, hence, dictates the share price change.

As for the optimization, the team considers that more sophisticated financial objectives and constraints should be added. By doing this, the optimization model can be made to be more robust and powerful especially in a dynamic financial market. In addition, we can further segment investors with more specific characteristics other than risk tolerance. A better segmentation method can help the team to find the investor type that our model can serve the best and also add more customizable parameters to improve usability.

7.0 Reference

[1]Ca.finance.yahoo.com. (2019). *Yahoo is now a part of Verizon Media*. [Online] Available at: https://ca.finance.yahoo.com

[2]"2019 Pricings | IPOScoop", *Iposcoop.com*, 2019. [Online]. Available: https://www.iposcoop.com/current-year-pricings/.

[3] The Balance. (2019). The 8 Best Investment Banks of 2019. [Online] Available at: https://www.thebalance.com/best-investment-banks-4177807 .

[4] Z. Tan, Z. Yan, and G. Zhu, “Stock selection with random forest: An exploitation of excess return in the Chinese stock market,” Heliyon, 17-Aug-2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844019359705.

8.0 Appendices

Appendix A - Correlation analysis specific to the target price

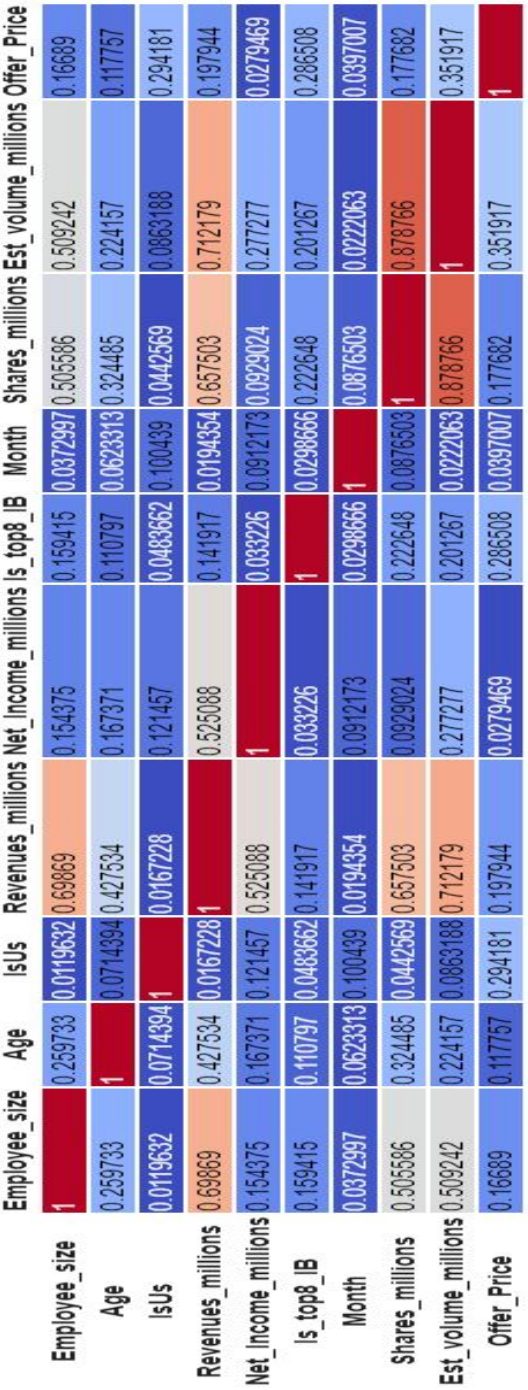
Employee_size 0.044811
Age 0.019082
IsUs 0.249181
Revenues_millions 0.074786
Net_Income_millions 0.038510
Is_top8_IB 0.261730
Month 0.042795
Shares_millions 0.100827
Est_volume_millions 0.219530
Offer_Price 0.816990
First_Day_Close 1.000000
Name: First_Day_Close, dtype: float64

Appendix B - Correlation analysis after one hot encoding specific to the target variable

First_Day_Close 1.000000
Offer_Price 0.816990
Is_top8_IB 0.261730
IsUs 0.249181
Est_volume_millions 0.219530
Technology 0.195540
Healthcare 0.102443
Shares_millions 0.100827
NASDAQ 0.092964
NYSE 0.092964
Consumer Goods 0.085377
Revenues_millions 0.074786
Telecommunications 0.061007
Financials 0.056185
Health Care 0.051852
Employee_size 0.044811
Month 0.042795
Net_Income_millions 0.038510
Industrials 0.032887
Energy 0.026371
Age 0.019082
Consumer Services 0.003254
Oil & Gas 0.002090
Name: First_Day_Close, dtype: float64

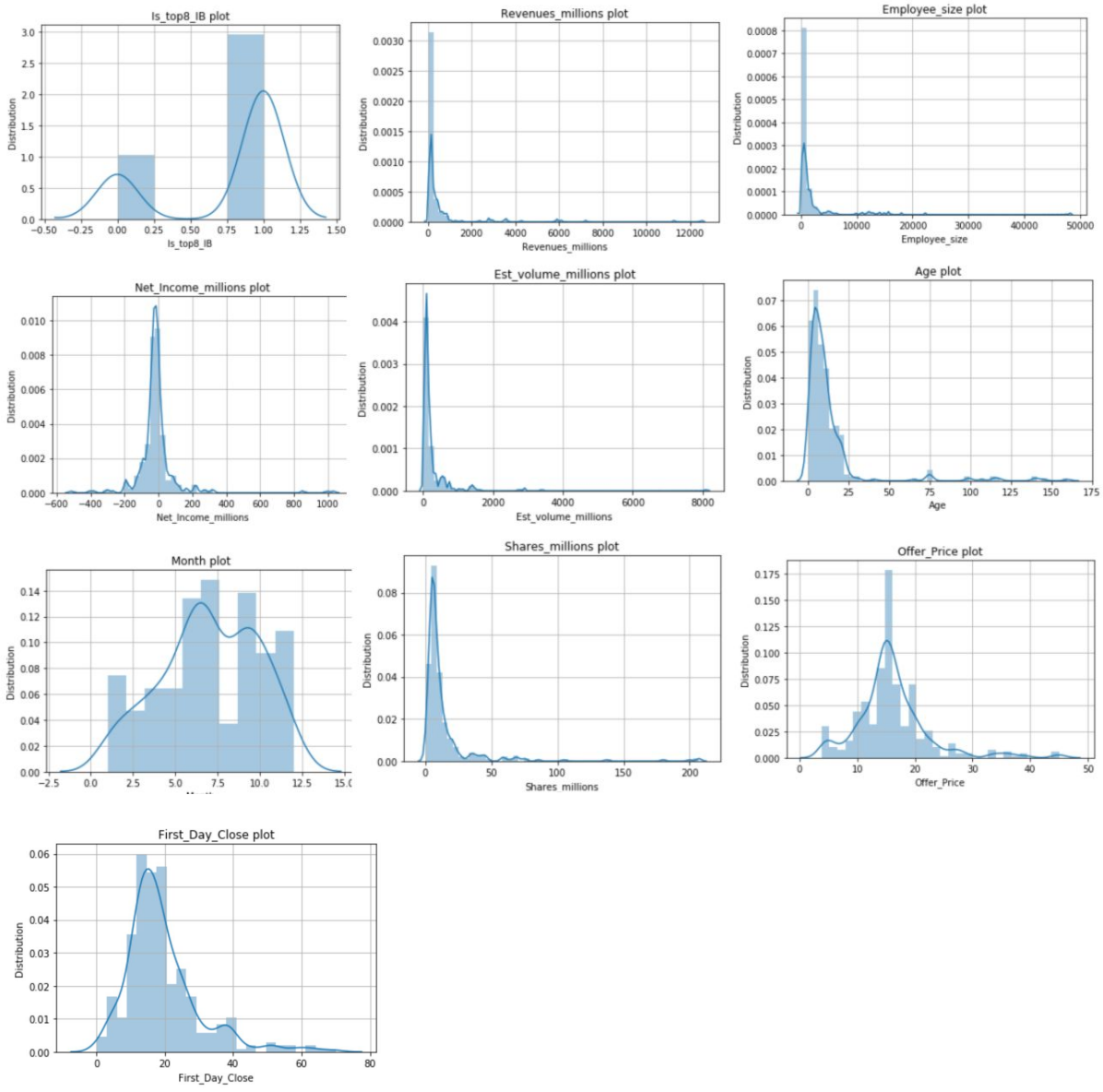
Appendix C - Correlation among all features

Figure 5 - Heat map showing the correlations between all the features



Appendix D - Features' distribution plots

Figure 6 - Distribution diagram of all the features



Appendix E - Outliers Analysis

The outlier analysis is conducted by plotting the scatter plot for the target value and other features, there is no extremely obvious outlier is spotted, then the team decided not to delete any data points. For example, Figure 7 below is a plot of the variable "Offer Price" and the target variable "First Day Close Price" using a joint plot function to check for outliers.

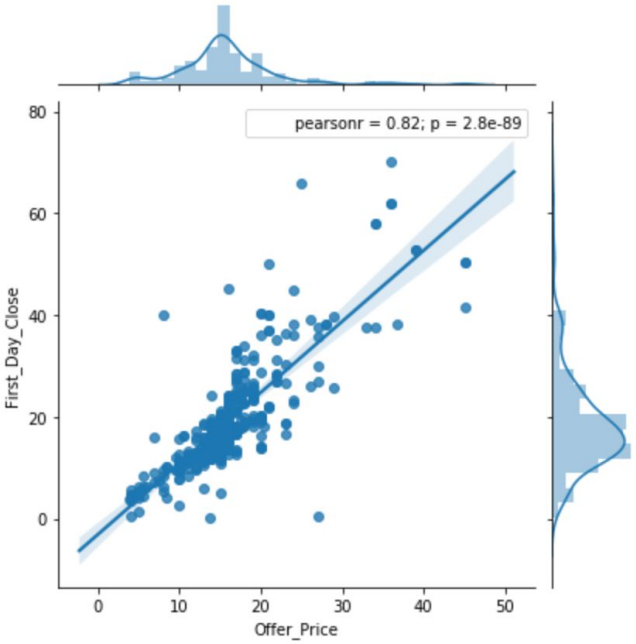


Figure 7 - Offer Price vs First Day Close Price