Group 24
Crystal Hu (jh2346), Jacob Dentes (jmd477), Michael Cao (yc849), Sanjana Kasetti (sk2465)

An Analysis of Shark Tank: Investment Behaviors

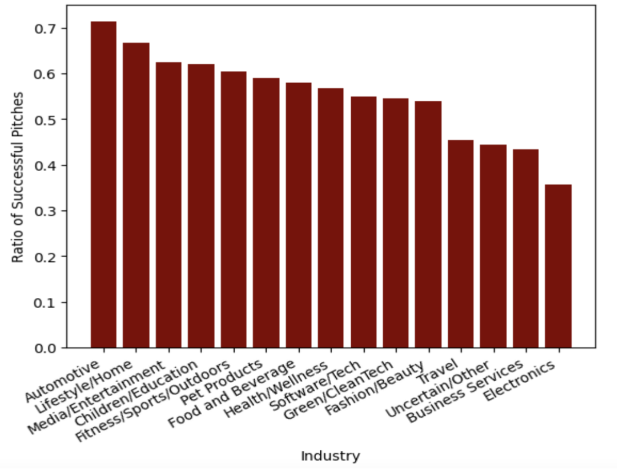
Abstract

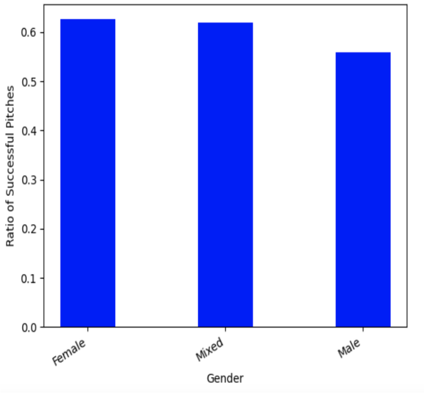
Angel investors undertake the critical task of accurately assessing the potential of a small business, while entrepreneurs are responsible for effectively presenting their venture to attract the best offers from investors. Shark Tank, a captivating TV show where budding entrepreneurs pitch their business ideas to a panel of investors known as sharks, provides a fascinating backdrop for this data analysis project. In this paper, we employ several statistical methods such as logistic regression and multiple linear regression to analyze the investment tendencies of each shark, shedding light on their unique strategies and preferences while highlighting the complexity of investment decision-making. Our research could be used by entrepreneurs who are interested in determining which aspects they should focus on most to maximize their chance of receiving an offer from a potential investor. It may also inform angel investors about the factors most likely to optimize their decision-making processes.

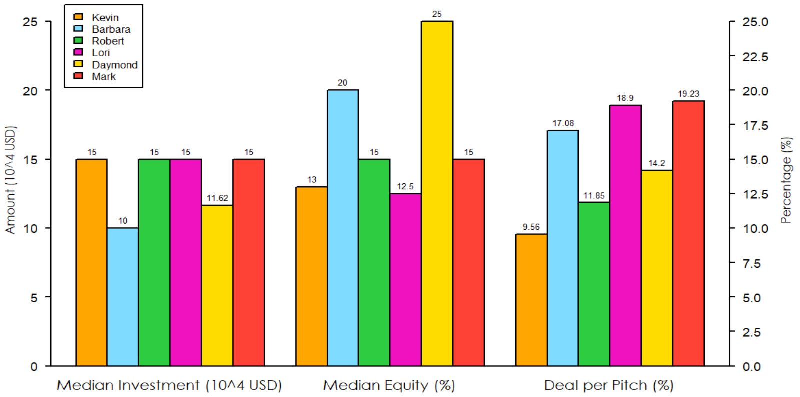
Introduction

i) Dataset Description

The dataset initially contained 50 columns and 1097 rows and it was arranged in wide format, with each row corresponding to a unique pitch on Shark Tank over the course of 14 seasons. The response variable in this dataset for the logistic model is 'GotDeal', an indicator variable which 1 corresponds to a successful pitch while 0 corresponds to an unsuccessful one, and 'TotalDealAmount' for the linear model, which is a numerical variable determined by the ratio between the deal amount and deal equity given that a pitch is successful. There are also 6 columns each containing an indicator variable indicating if they are present for the pitch, which we used to subset the dataset into separate datasets in preparation for our further analysis focused on each shark. Prior to analyzing the data, we have also omitted columns where most entries are incomplete or erroneous, such as 'PitcherAge' and 'PitcherCity', where some entries in 'PitcherGender' were entered manually based on public information to make up for the missing entries. This was done to ensure that the values in each column are well-defined.







Some important covariates we considered prior to model fitting include two categorical variables and three numerical variables, where a selected few can be visualized by the figures above. The two numerical variables are ‘Industry’, which is a categorical variable with 15 factor levels to describe the pitcher’s industry, and ‘PitchersGender’, ‘SeasonNumber’, which has 3 factor levels - male, female, and mixed team. Both categorical variables have been dummy coded to be meaningful to our data analysis. The numerical variables include ‘PitchNumber’, which accounts for the shift in investment tendencies over time, ‘OriginalAskAmount’ and ‘OriginalOfferedEquity’, which constitute the offer proposed by the entrepreneur prior to giving their pitch.

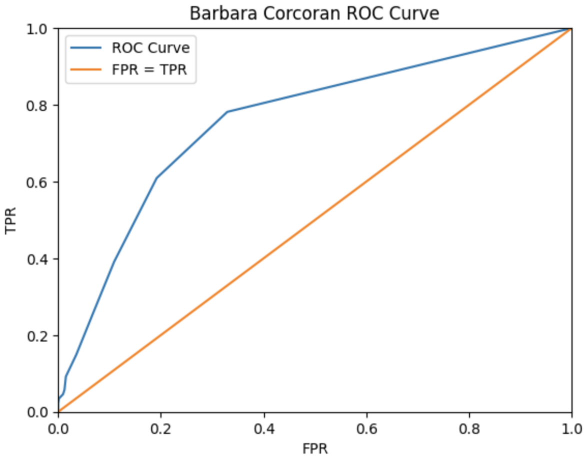
ii) Research Questions and Methods

Our primary focus revolves around addressing two overarching questions, where each would be approached through an overall group investigation, followed by a more in-depth analysis of the investment behaviors exhibited by individual sharks. First, we will employ logistic regression to predict whether a pitcher will obtain a deal based on relevant information about their business. Next, we will subset the data to include only those who successfully secured a deal, then employ multiple linear regression to model the final deal amount. Each model described above would be constructed using a train-test split approach to accurately assess the model fit and avoid overfitting. Specifically, 75% of the rows will be randomly assigned to the training set, while the remaining 25% will comprise the test set.

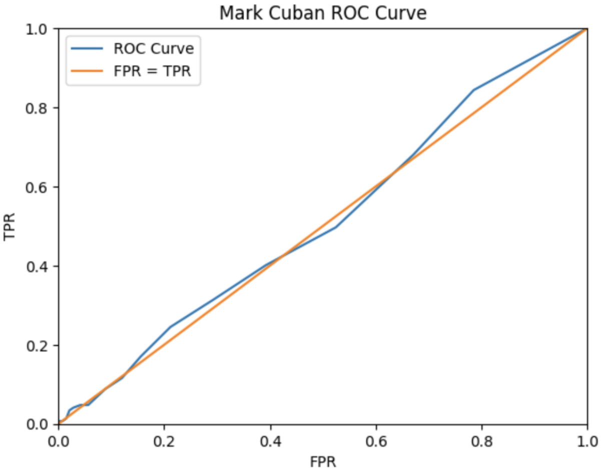
Results

i) Logistic Regression: Per-Shark Deal Prediction

We used logistic regression to predict whether or not a given pitch would close a deal on the show. We fit a model for each shark to predict whether or not that shark would make a deal with a given pitcher. The predictors used in the model were: the pitch number, whether or not the business had multiple entrepreneurs, the amount of money that the pitcher originally asked for, the amount of equity that the pitcher offered, whether or not each shark was present, the pitcher's gender, and the business' industry. Categorical variables were handled using the Statsmodels formula API for Python. Each model was trained on a subset of 75% of the data and we created ROC curves plotting the true positive rate versus false positive rate for the remaining 25% of the data. We used the ROC curves to evaluate how successfully the model could predict investment behavior.



The logistic regression model produced promising results for some sharks. For example, using the model to predict whether or not Barbara Corcoran would invest produced the ROC curve in "Barbara Corcoran ROC Curve." An orange diagonal line is also plotted to represent the accuracy of a random model. The signed area between the ROC curve and the forty-five degree line is 0.258. This suggests that the information in the dataset is particularly effective for predicting whether or not Corcoran will invest in a given pitcher. Examining the model's summary gives increased insight into this observation. Many covariates were significant at the 95% level. For example, the model has coefficients of -0.59 and -1.04 for if the pitchers are male or mixed gender, respectively. The p-values for these coefficients were 0.023 and 0.009, respectively, suggesting that they are significant at the 95% confidence level. This suggests that Corcoran is most likely to invest in female pitchers and least likely to invest in a group of pitchers with mixed genders. These statistically significant covariates are predictors in our model, which gives the model improved accuracy.



Conversely, the model was a poor predictor for some other sharks' behavior. For example, using the model to predict whether or not Mark Cuban would invest produced the ROC curve in "Mark Cuban ROC Curve". The signed area between this ROC curve and the forty-five degree line is only 0.0128, which is much smaller than the area in Barbara's case. This suggests that the model does not predict much better than random chance. The model summary gives insight into why it performed so poorly. According to the computed p-values, the only statistically significant predictors at the 95% level were whether or not the business was in the pet product industry and whether or not Mark Cuban was present for the pitch. This suggests that most of the predictors present in the data set are relatively weak indicators for whether or not Mark Cuban invests, and he likely relies more on other factors when deciding whether or not to invest. The table on the right shows the signed area between the generated ROC curve and the forty-five degree line for each shark. Based on the signed area for each shark, the information provided in the dataset provides strong insight into the investment behavior of some sharks, like Barbara Corcoran or Lori Greiner, but does not provide much insight into the investment behavior of other sharks, like Kevin O'Leary or Mark Cuban.

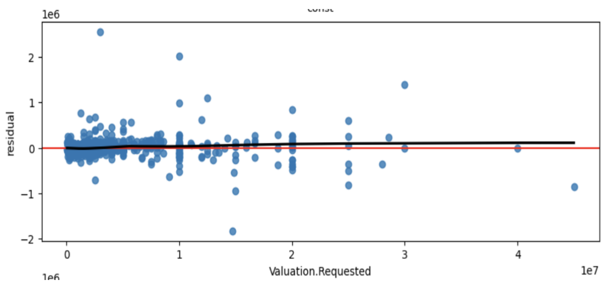
Shark	Signed Area
Barbara Corcoran	0.258
Lori Greiner	0.184
Daymond John	0.169
Robert Herjavec	0.103
Kevin O'Leary	0.0421
Mark Cuban	0.0128

Overall, our findings suggest that deciding whether or not to invest in a given entrepreneur is a complicated issue, and different investors likely weigh information differently when making a decision about how to use their money. The model likely worked well on sharks that consider the model's covariates important when making their investment decision. Conversely, it likely worked poorly on sharks that place more importance on factors outside the dataset. This model could be helpful for several groups of people. It may be helpful for the producers of Shark Tank when deciding on entrepreneurs to put on the show because it can give insight into the probability that a deal will go through. It also may be helpful for prospective contestants to the show. A contestant could use the model to see what their weaknesses are prior to a pitch so that they can improve their chances of getting a deal from the sharks.

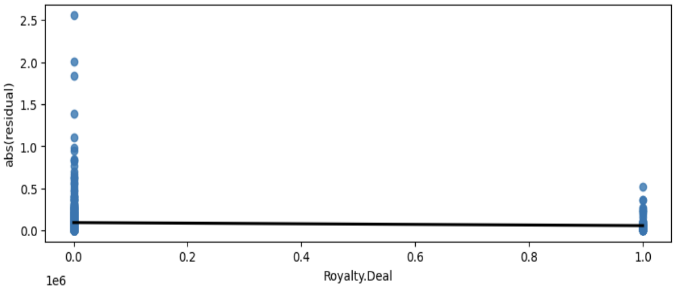
ii) Testing Linear Assumptions

We tested assumptions of linear regression using the dependent variable of Total Deal Amount against the predictors: ValuationRequested, RoyaltyDeal, DealValuation, Numberofsharkindeal, OriginalAskAmount(the guest asked investment), OriginalOfferedEquity (the guest asked equity), GotDeal (represents if a guest got a deal), SeasonNumber, PitchNumber (the guest appearance number), Loan, MultipleEntrepreneurs (whether multiple people were guests), BarbaraCorcoranPresent, MarkCubanPresent, LoriGreinerPresent, RobertHerjavecPresent, DaymondJohnPresent, KevinO'LearyPresent (all the present predictors represent whether each shark was at this pitch), Industry (what industry was the guest's product in).

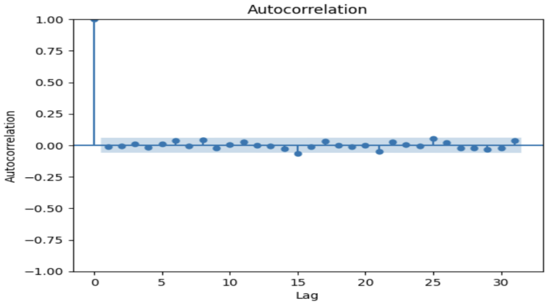
The first assumption we tested was whether the model is linear in parameters. The predictors that are linear are the ones where the residual line is at y = 0, so we plotted y = 0 in red and the residuals of each predictor to compare the lines. As we can see in the



graph on the right, the model is linear in parameters with predictors RoyaltyDeal, NumberofSharksinDeal, OriginalOfferedEquity, GotDeal, SeasonNumber, PitchNumber, Loan, MultipleEntrepreneurs, sharks Barbara, Lori, Robert, Daymond, and Kevin are present, and industries: BusinessServices, Children/Education, Electronics, Fashion/Beauty, Fitness/Sports/Outdoors, Food/Beverage, Lifestyle/Home, Pet Products, Software/Tech. We list the predictors whose lowess line is close to y = 0 with little variation. We only show one graph as an example of what the residual plot looks like. The rest of the graphs of each predictors residual are displayed in the appendix.

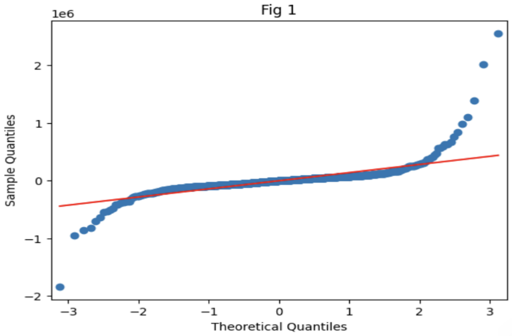


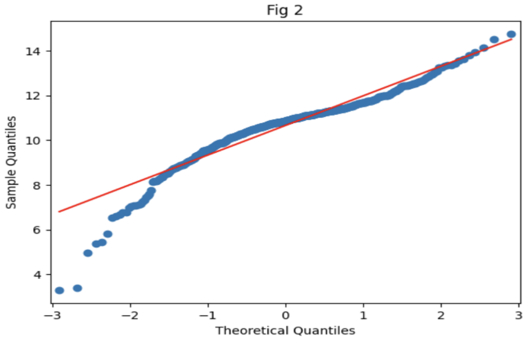
The second assumption we tested was whether the errors have constant variance. This is determined by plotting the square residuals for each predictor and observing whether there is a horizontal lowess line. Thus, as we can see on the graph to the right, the predictors have errors with approximately constant variance are: RoyaltyDeal, NumberofSharksinDeal, OriginalofferedEquity, GotDeal, SeasonNumber, PitchNumber, Loan, MultipleEntrepreneurs, BarbaraCorcoranPresent, LoriGreinerPresent, RobertHerjavecPresent, DaymondJohnPresent, and Industries: Business Services, Children/Education, Electronics, Fashion/Beauty, Fitness/Sports/Outdoors, Food/Beverage, Green/Clean Tech, Lifestyle/Home, Media/Entertainment, Pet Products, Software/Tech. We only show one graph as an example of what the squared residual plot looks like. The rest of the graphs of each predictors squared residual plot are displayed in the appendix.



The third assumption we tested was whether the errors are independent. We determined this by plotting the autocorrelation plot of the residuals, which is on the left. Since the lag vs. autocorrelation plot had only one large peak at x = 0, we could conclude that the errors are independent.

Lastly, we tested whether the errors are normally distributed. We determined this by plotting a QQ plot of the residuals. From the Fig 1 plot (below) we can see that errors are heavy tailed and are not normal. Even though the errors are not normal, we can make them closer to a normal distribution by taking the natural log of the residuals, which produces the Fig 2 plot (below) that is closer to a normal distribution.



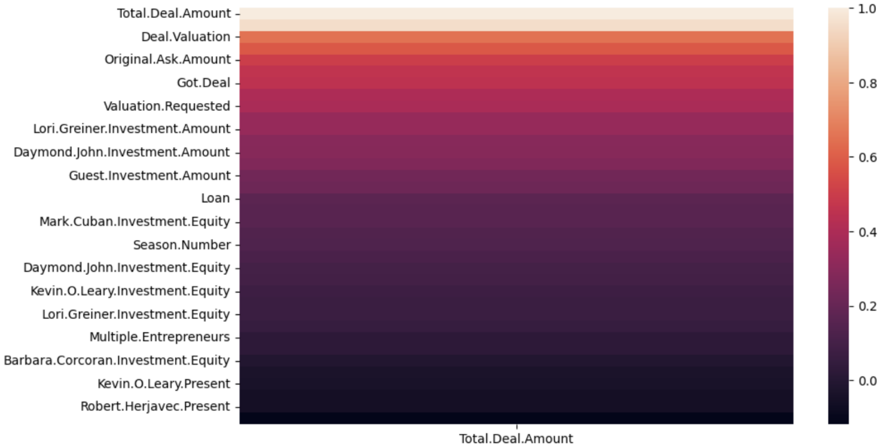


Through these plots we can see that Total Deal Amount is linear with respect to Industries: Business Services, Children/Education, Electronics, Fashion/Beauty, Fitness/Sports/Outdoors, Food/Beverage, Lifestyle/Home, Pet Products, and Software/Tech. In addition, TotalDealAmount is linear with predictors RoyaltyDeal, NumberofSharksinDeal, OriginalOfferedEquity, GotDeal, SeasonNumber, Pitch Number, Loan, and MultipleEntrepreneurs. While there might not be much to infer knowing that SeasonNumber, PitchNumber and GotDeal is linear with regards to TotalDealAmount, we can perhaps predict TotalDealAmount using how much the OriginalOfferedEquity, how much the guests asked for in loan, and the number of sharks in deal.

iii) Linear Regression: Deal Investment Amount Prediction

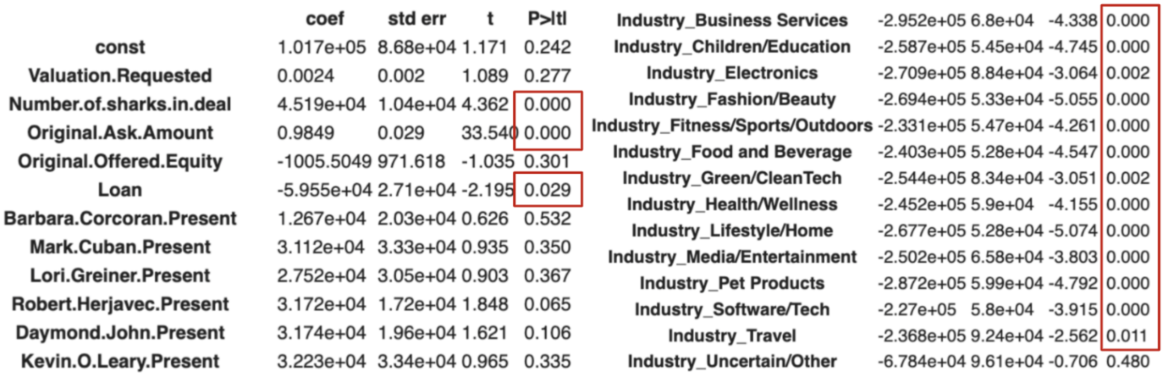
We built a linear regression model to analyze and predict the total investment amount for a successful pitch based on a set of predictor variables. Through the analysis, we aimed to determine how each predictor variable associated with the total investment deal value by quantifying the magnitude and direction of associations using linear regression models.

First of all, we conducted a correlation analysis on the total investment amount in the shark tank dataset for all the variables. The purpose was to examine the pairwise associations and significance of correlation between each variable in the dataframe and total deal investment amount.



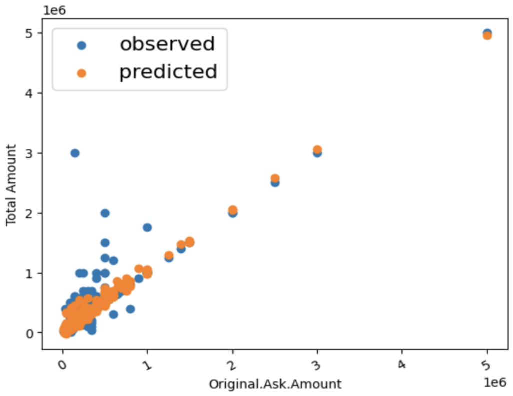
The correlation analysis presented several significant correlations between the variables in the dataframe, such as OriginalAskAmount (0.505), and TotalDealAmount shown in the correlation heatmap. Among all, the strongest positive correlation was observed between OriginalAskAmount and TotalDealAmount, with a correlation coefficient of 0.505158, which suggests that as OriginalAskAmount increases, TotalDealAmount also tends to increase. In contrast, the strongest negative correlation was observed between OriginalOfferedEquity and TotalDealAmount, with a correlation coefficient of -0.119, suggesting that as OriginalOfferedEquity increases, TotalDealAmount tends to decrease. The results of correlation analysis shed light on selecting variables in the dataframe as features to observe in our linear regression model.

After considerations, the predictors chosen are ValuationRequested, NumberOfSharksInDeal, OriginalAskAmoung, OriginalOfferedEquity, Loan, Industry, and whether each shark was present, where Industry is a categorical variable. Our model had an R-squared value of 0.819, demonstrating that the predictor variables in our model explained around 81.9% of the variance in the total investment deal value. Coefficients of each predictor demonstrate when all other variables are held constant, an one-unit increase in each predictor can cause change in the total investment deal value. The p-values for each coefficient were also presented: p-value less than 0.05 indicates a statistical significance of coefficient. Among 25 variables (11 predictor variables + one dummy variable for each industry), there were 16 variables' p-value < 0.05, showing that our model was statistically significant and that some of the predictor variables had a significant relationship with the total investment value.



	coef	std err	t	P>ltl
const	1.017e+05	8.68e+04	1.171	0.242
Valuation.Requested	0.0024	0.002	1.089	0.277
Number.of.sharks.in.deal	4.519e+04	1.04e+04	4.362	0.000
Original.Ask.Amount	0.9849	0.029	33.540	0.000
Original.Offered.Equity	-1005.5049	971.618	-1.035	0.301
Loan	-5.955e+04	2.71e+04	-2.195	0.029
Barbara.Corcoran.Present	1.267e+04	2.03e+04	0.626	0.532
Mark.Cuban.Present	3.112e+04	3.33e+04	0.935	0.350
Lori.Greiner.Present	2.752e+04	3.05e+04	0.903	0.367
Robert.Herjavec.Present	3.172e+04	1.72e+04	1.848	0.065
Daymond.John.Present	3.174e+04	1.96e+04	1.621	0.106
Kevin.O.Leary.Present	3.223e+04	3.34e+04	0.965	0.335
Industry_Business Services	-2.952e+05	6.8e+04	-4.338	0.000
Industry_Children/Education	-2.587e+05	5.45e+04	-4.745	0.000
Industry_Electronics	-2.709e+05	8.84e+04	-3.064	0.002
Industry_Fashion/Beauty	-2.694e+05	5.33e+04	-5.055	0.000
Industry_Fitness/Sports/Outdoors	-2.331e+05	5.47e+04	-4.261	0.000
Industry_Food and Beverage	-2.403e+05	5.28e+04	-4.547	0.000
Industry_Green/CleanTech	-2.544e+05	8.34e+04	-3.051	0.002
Industry_Health/Wellness	-2.452e+05	5.9e+04	-4.155	0.000
Industry_Lifestyle/Home	-2.677e+05	5.28e+04	-5.074	0.000
Industry_Media/Entertainment	-2.502e+05	6.58e+04	-3.803	0.000
Industry_Pet Products	-2.872e+05	5.99e+04	-4.792	0.000
Industry_Software/Tech	-2.27e+05	5.8e+04	-3.915	0.000
Industry_Travel	-2.368e+05	9.24e+04	-2.562	0.011
Industry_Uncertain/Other	-6.784e+04	9.61e+04	-0.706	0.480

Our model results indicated that some predictor variables like OriginalAskAmount, with a coefficient of 0.985, had a significant positive relationship with the total investment deal value. As OriginalAskAmount increases by one unit, the total investment deal value increases by 0.985 unit. Then, we use the predictor variables to predict total deal amounts. We plot a scatter plot with OriginalAskAmount on x-axis and both predicted and observed TotalDealAmount on y-axis. The scatter plot indicated a high alignment between predicted and observed total deal amount values. In contrast, Loan (coefficient=-6.199e+04) had a significant negative relationship with the total investment deal value. It demonstrated that when Loan is offered, the total investment deal value decreases by -6.199e+04 unit compared to the case when Loan is not offered. Some other predictor variables did not have significant relationships with the total investment deal value, but they gave an interesting overview of the dataset and linear model, such as ValuationRequested and OriginalOfferedEquity.



Conclusion

In this project, we utilized logistic regression to assess the likelihood of Shark Tank pitchers securing deals based on relevant information about their businesses. Our research findings shed light on the intricate nature of investment decision-making, with investors assigning varying weights to different factors. While the model proved effective for sharks who prioritize the model's covariates, such as gender and industry, it yielded less favorable results for those who consider external factors, adopt a more holistic investment approach, or possess a versatile portfolio. Notably, the significance of each predictor varies significantly across different models.

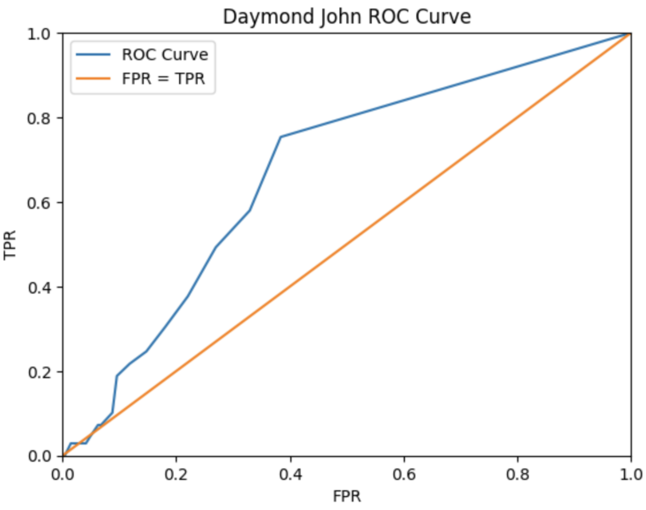
Furthermore, employing multiple linear regression, we sought to predict the total deal amount for successful pitches. Our analysis uncovered significant correlations. The original ask amount exhibited the most substantial positive correlation with the total deal amount, while offering a loan displayed a significant negative relationship. The industry in which the business operates emerged as one of the most influential predictors, and a strong positive correlation was observed when multiple sharks participated in a deal. Ultimately, our linear model achieved an impressive R-squared value of 0.819, indicating an overall well-fitted model.

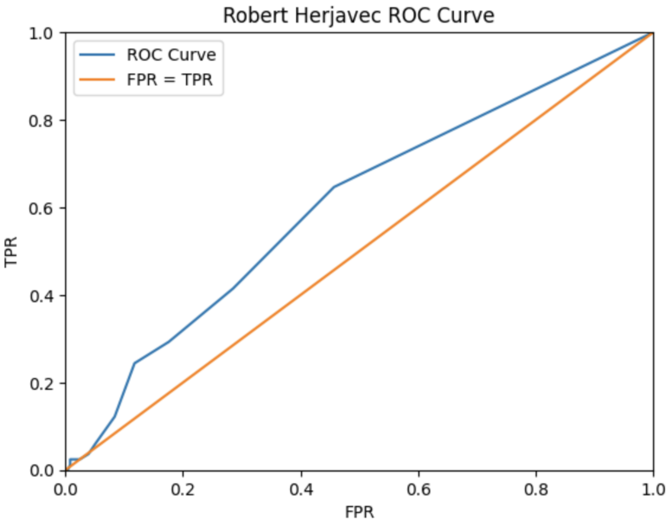
However, it is important to acknowledge the limitations of our study. The analysis relies on data solely from the Shark Tank TV show, which may not fully capture the intricacies of real-world investment scenarios. Additionally, the predictive power of the models may vary among different sharks, as they cannot guarantee accurate predictions in all cases. Nevertheless, our research contributes to understanding investment behaviors and offers practical implications for entrepreneurs and angel investors alike.

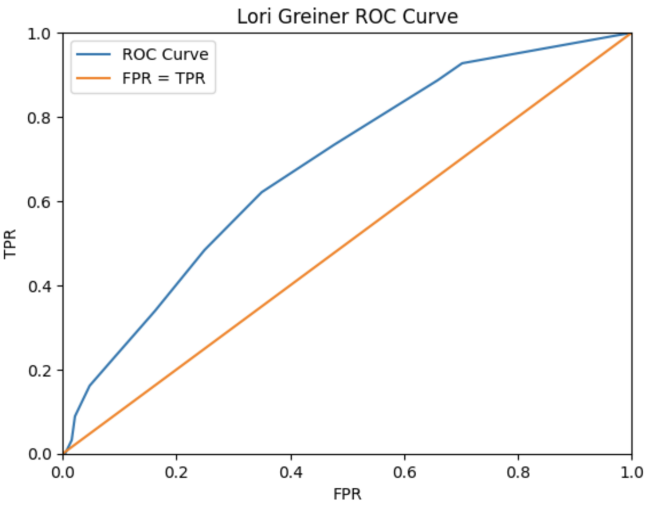
Overall, our findings underscore the complexity of investment decision-making and emphasize the importance of considering individual investor preferences and strategies. The models developed in this report provide valuable insights for entrepreneurs aiming to maximize their chances of securing investment offers and for the show’s producers when selecting entrepreneurs. Furthermore, our analysis offers a glimpse into the factors influencing investment decisions and establishes a foundation for future research in this domain.

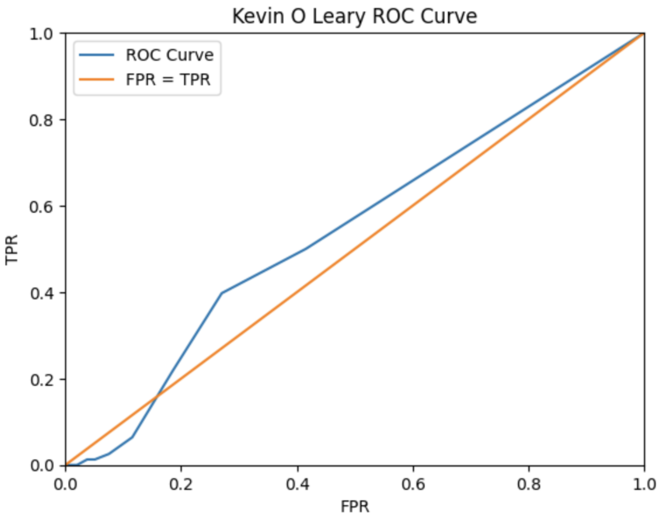
Appendix

i) Additional Plots for Logistic Regression



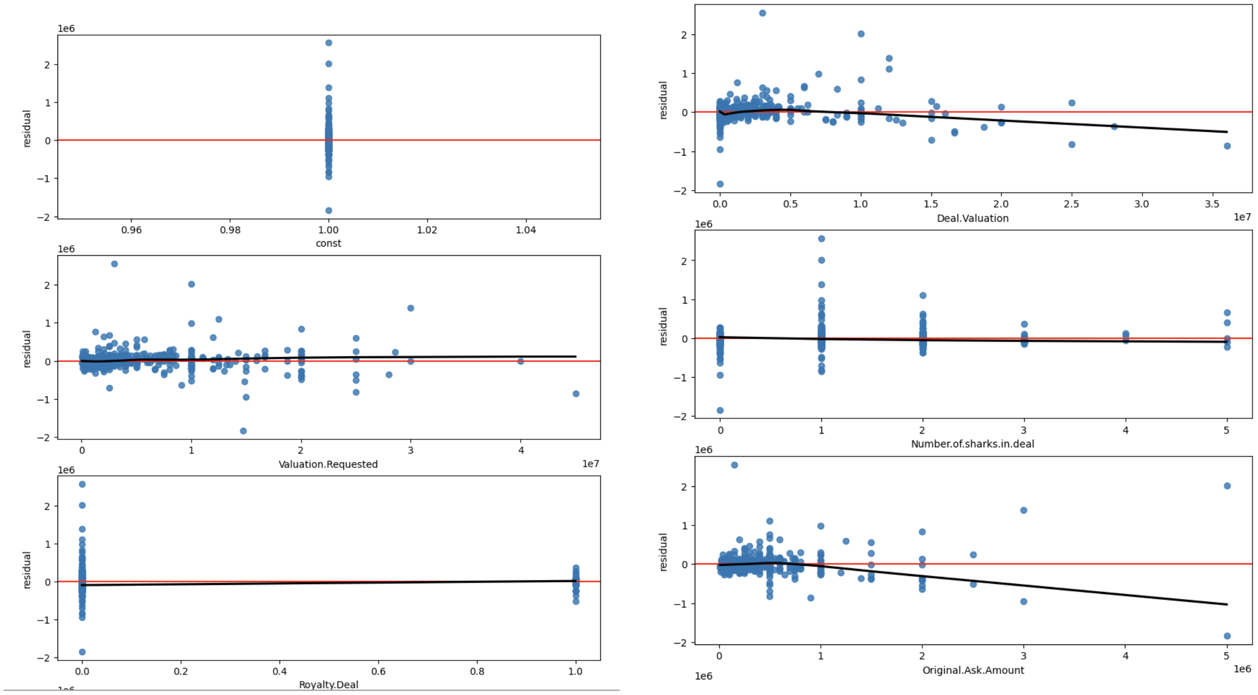


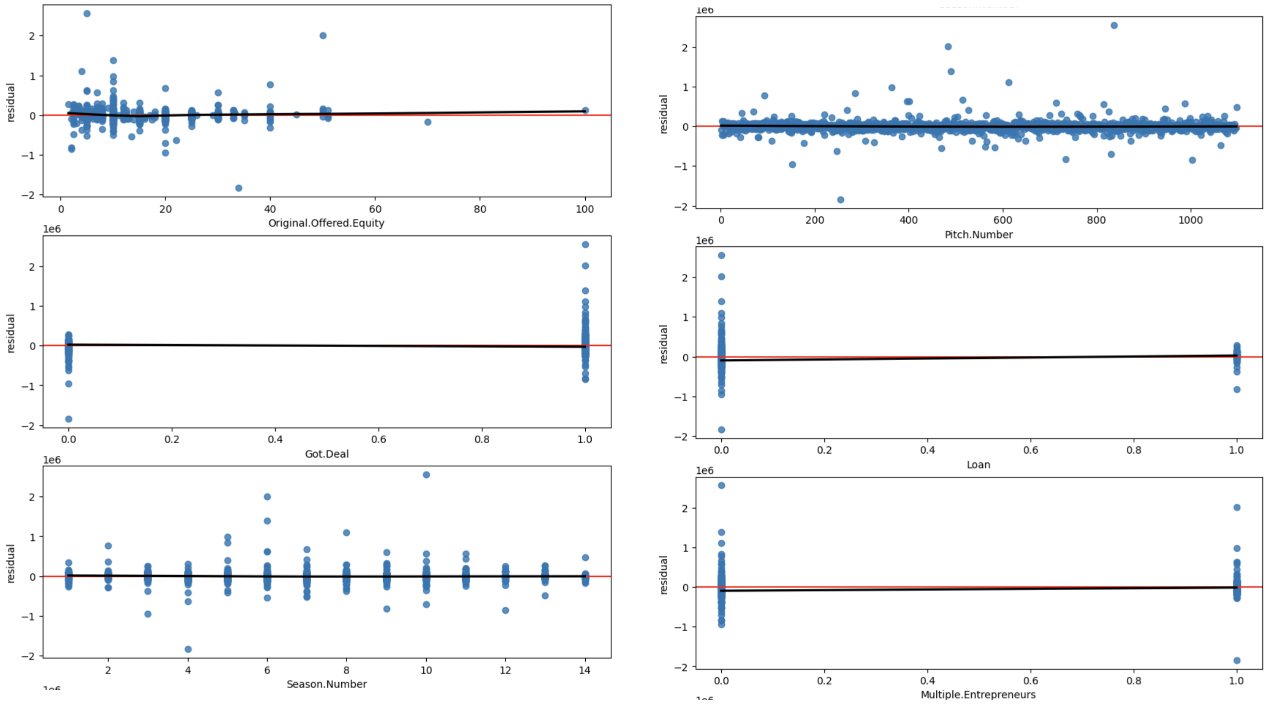


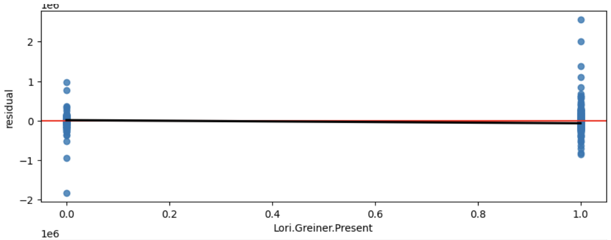


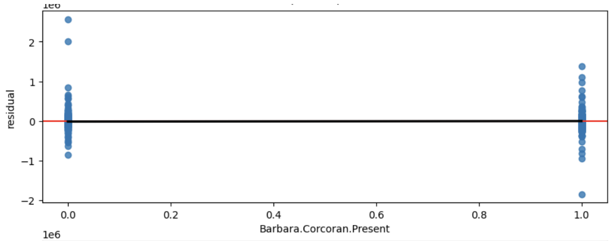
ii) Additional Plots for Assumptions of Linear Regression

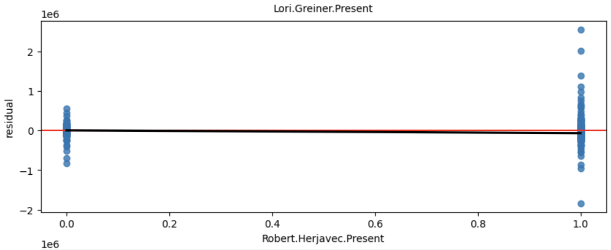
Residual Plots for Assumption 1

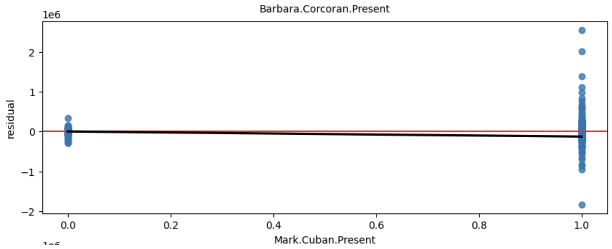


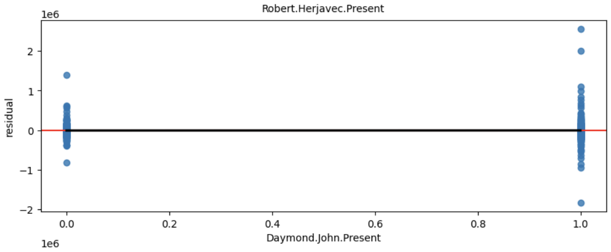


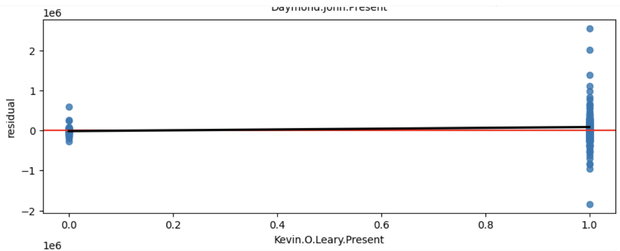

1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Lori.Greiner.Present

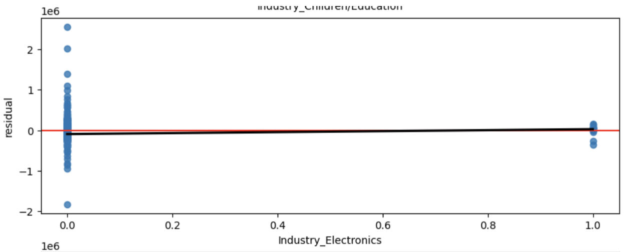

1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Barbara.Corcoran.Present

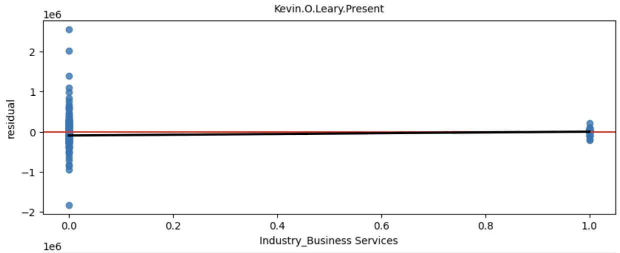

1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Robert.Herjavec.Present

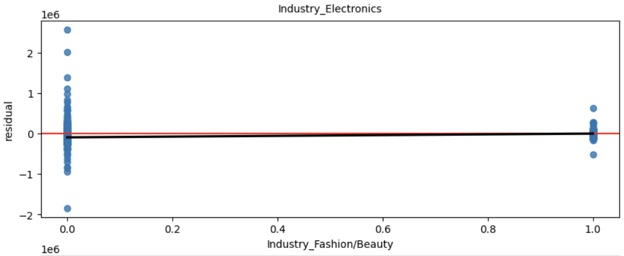

1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Mark.Cuban.Present

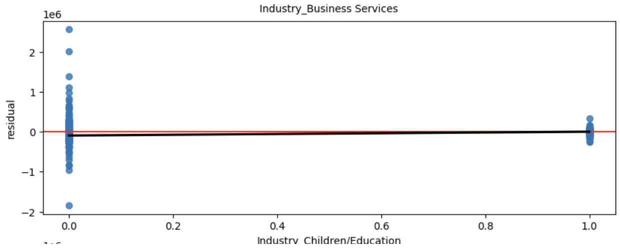

1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Daymond.John.Present
1e6

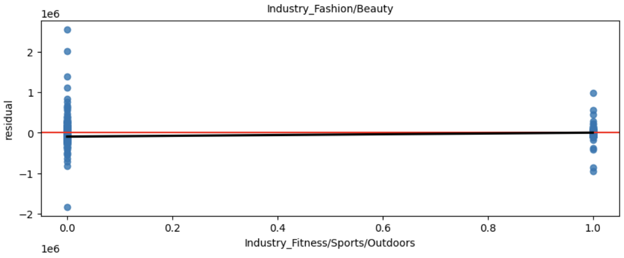

Daymond.John.Present
1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Kevin.O.Leary.Present

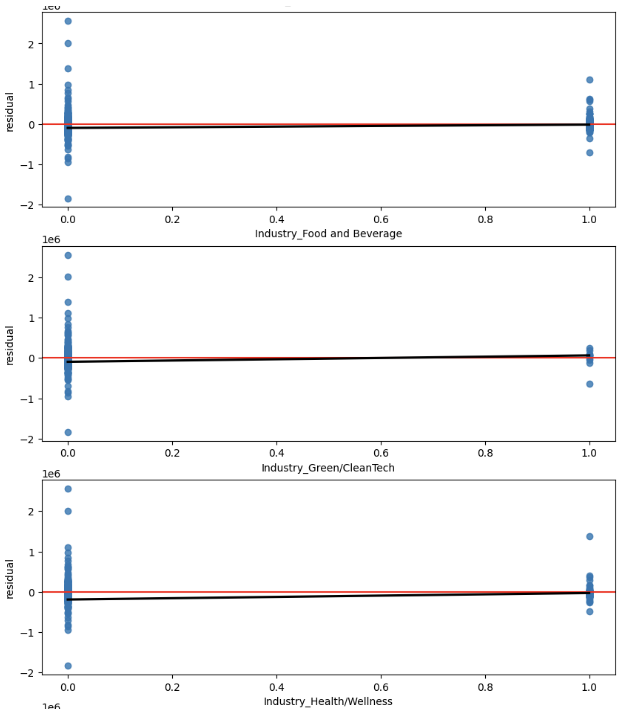

Industry_Children/Education
1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Electronics

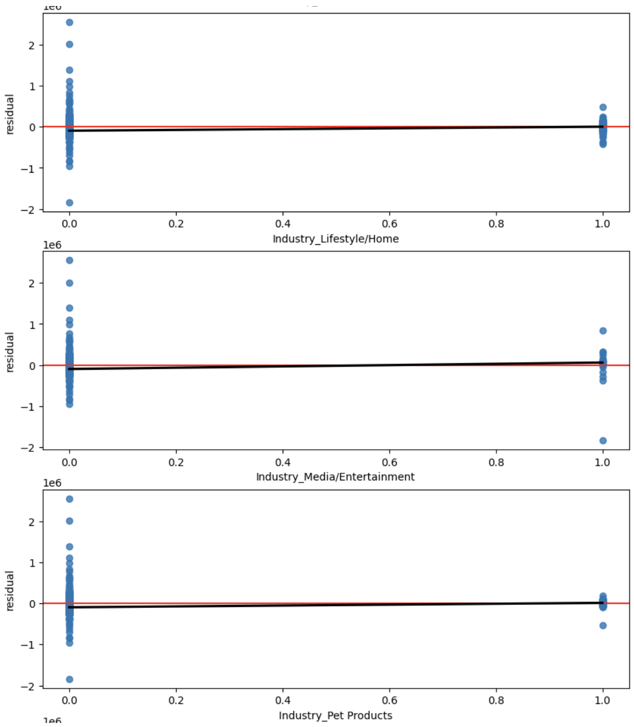

1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Business Services

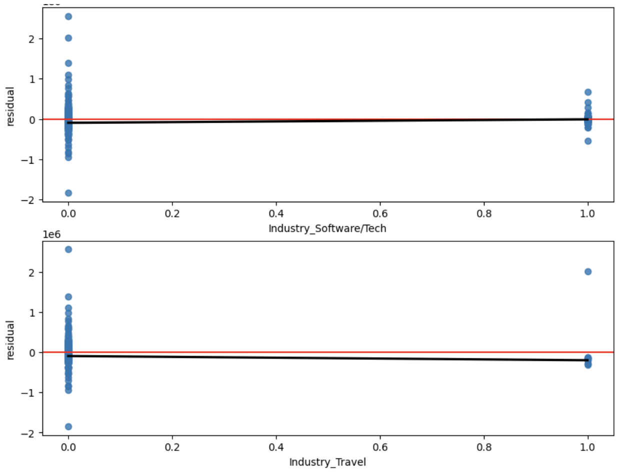

1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Fashion/Beauty


1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Children/Education

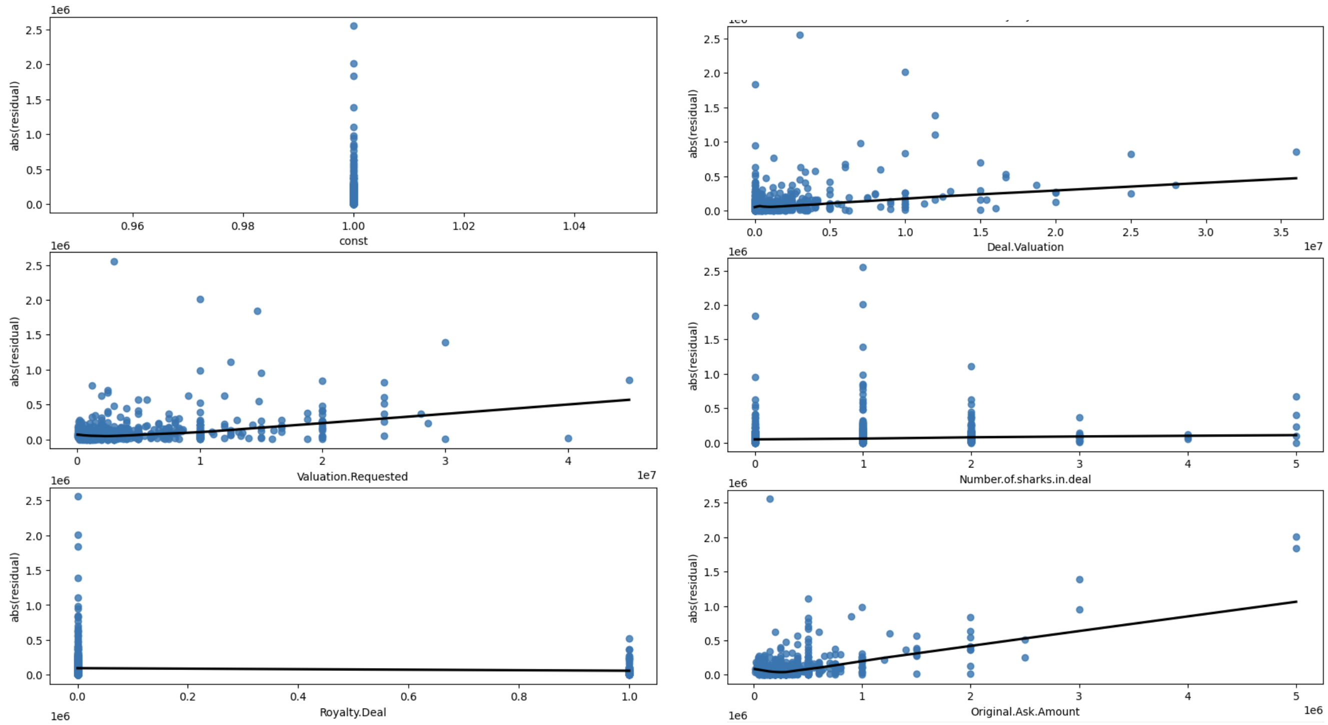

1e6
2
1
0
−1
−2
residual
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Fitness/Sports/Outdoors
1e6

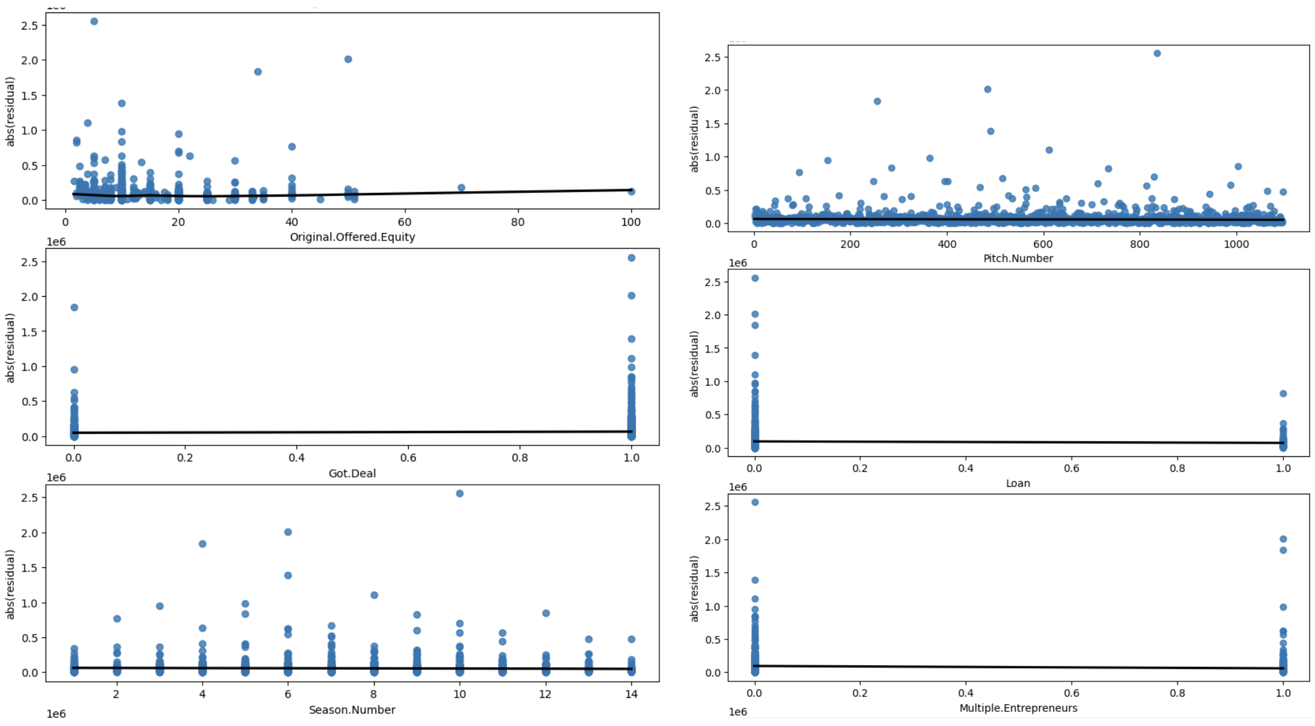

1e6
residual
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Food and Beverage
1e6
residual
Industry_Green/CleanTech
1e6
residual
Industry_Health/Wellness
1e6

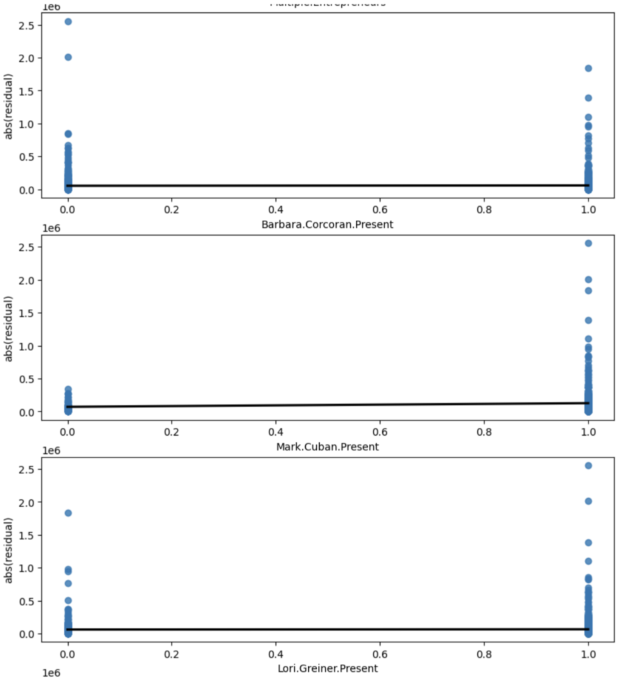

1e6
residual
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Lifestyle/Home
1e6
residual
Industry_Media/Entertainment
1e6
residual
Industry_Pet Products
1e6

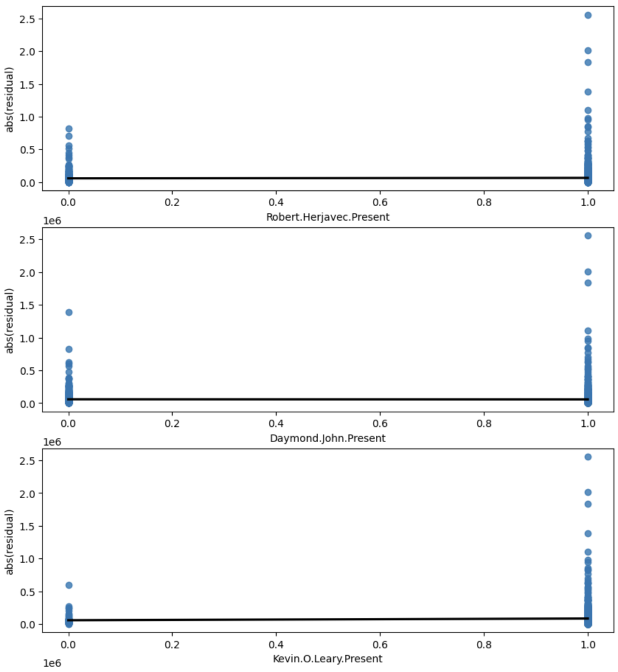

residual
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Software/Tech
1e6
residual
Industry_Travel

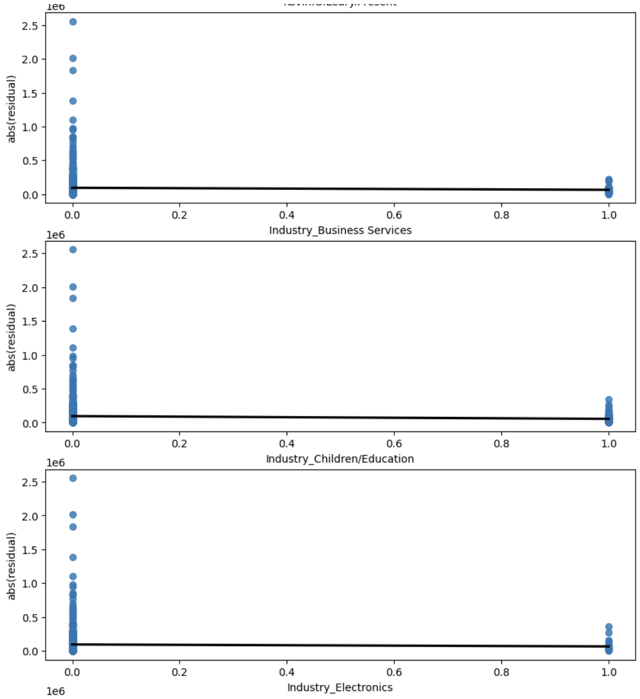
Square Residual Plots for Assumption 2

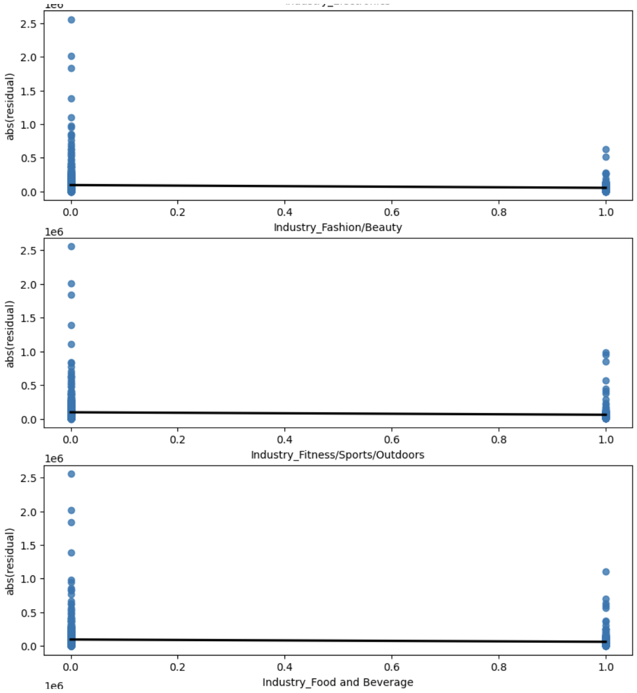


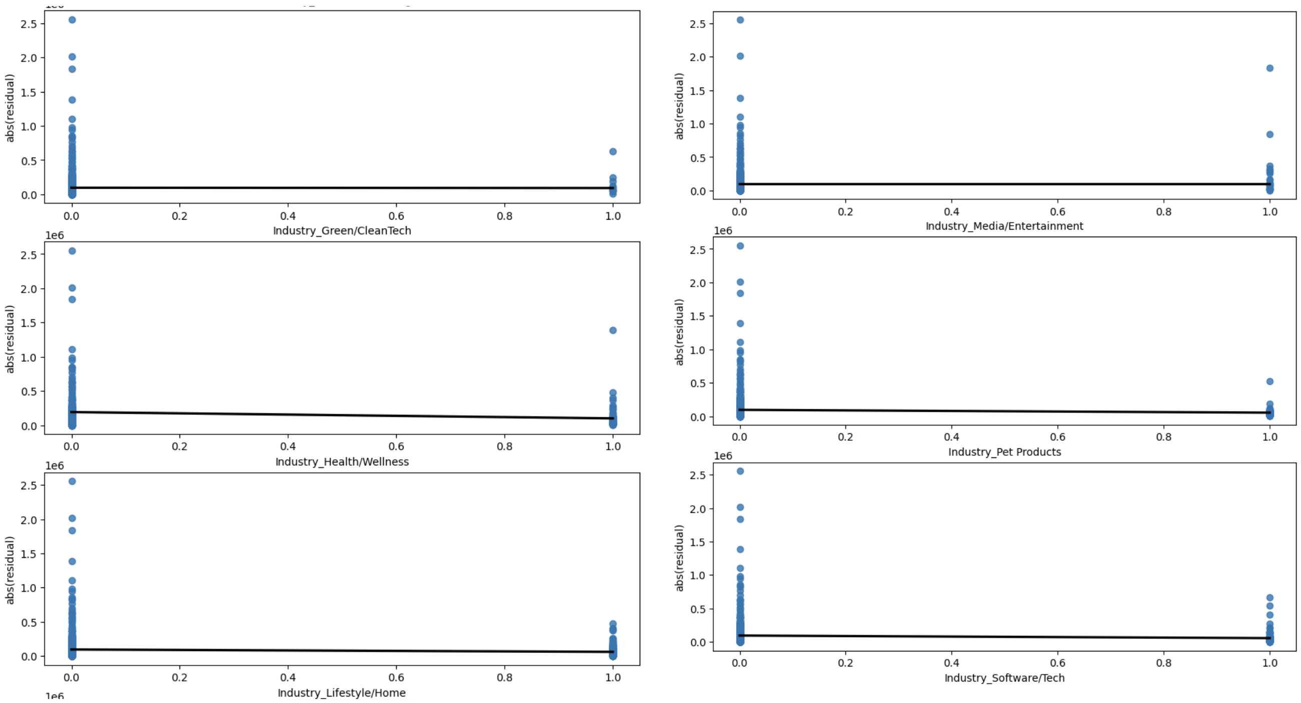



1e6
abs(residual)
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Barbara.Corcoran.Present
Mark.Cuban.Present
Lori.Greiner.Present


1e6
abs(residual)
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Robert.Herjavec.Present
Daymond.John.Present
Kevin.O.Leary.Present


1e6
abs(residual)
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Business Services
Industry_Children/Education
Industry_Electronics


1e6
abs(residual)
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Fashion/Beauty
Industry_Fitness/Sports/Outdoors
Industry_Food and Beverage


1e6
abs(residual)
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Industry_Green/CleanTech
Industry_Media/Entertainment
Industry_Health/Wellness
Industry_Pet Products
Industry_Lifestyle/Home
Industry_Software/Tech