SharkTank Deal Prediction: Dataset and Computational Model

Thomas Sherk
Department of Computer Science
University of Dayton
Dayton, United States
sherkt1@udayton.edu

Minh-Triet Tran
Department of Computer Science
University of Science
Ho Chi Minh City, Vietnam
tmtriet@fit.hcmus.edu.vn

Tam V. Nguyen
Department of Computer Science
University of Dayton
Dayton, United States
tamnguyen@udayton.edu

***Abstract*—SharkTank is a television show where start-ups pitch their idea to a panel of five investors (sharks) in hopes of striking a deal in the form of equity or royalties for money and other business perks. Since its inception, SharkTank has been a center of discussion and analysis for fans, statisticians, and business people alike in hopes of cracking the code to the start-up world and figuring out the formula for the next big 'thing'. Most of these discussions and analyses have come in the form of blogs, articles, and academic research. However, there has been a lack of complete datasets and application of computational models for further analysis. In this paper, we investigate factors that play into the SharkTank deal. To this end, we first collect a new dataset, SharkTank Deal Dataset (STDD), by combining data from multiple public sources. The dataset includes descriptive features of each start-up such as product category, team composition, valuation, equity offering, specific sharks that appear on that episode, and state origin. For the computational model, we propose a new computational model to predict whether a start-up strikes a deal with a shark. We conduct experiments to demonstrate the superiority of our model over the baselines.**

***Index Terms*—startup, deal prediction, dataset, computational model**

I. Introduction

Since its beginning, Shark Tank[1] has been a center of attention for fans which has led to the formation of popular blogs such as SharkTankBlog [15] and increased popularity amongst entrepreneurship majors and courses offered in business schools [24]. The reasons for this are multi-faceted beyond just its entertainment value: SharkTank is a semi-transparent form of what occurs at private venture capital firms (investment firms that specialize in startup investing). In each episode, the show documents the history of the startup (including the background of the founders), the 'pitch' to the panel of 'sharks', the offer the startup is willing to give (usually in the form of equity or royalties), and the bargaining phase between the startup and 'sharks'. If the viewers are so inclined, they can also follow the ongoing business after it appears on the show to see if the business deal was a success or not.

[1]https://abc.go.com/shows/shark-tank

Venture Capital (VC) firms traditionally take on greater risk than most investment funds, considering that the majority of startups fail [8] [26], but do so by investing in companies with large growth potential (partially due to industries with large growth potential, too) with the promise of equity amongst other benefits. It is the hope from the investors perspective that the return on investment for the successful companies to be some multiple of the original investment to compensate for other failed startups and the large amount of risk involved [27].

Startups that receive VC funding enjoy benefits from the VC other than just funding such as exposure, guidance from industry professionals, and connections to other VC funded companies. It is normal to hear sharks try to entice startups on the show with the promise of greater exposure (for example, Lori Greiner has a show on QVC where she can sell SharkTank products direct to consumer[2]) or other benefits such as access to production facilities or distribution channels with the intent of gaining a better deal for themselves or to outbid a shark without compromising their offer. But startups sometime require multiple rounds of funding as companies continue to grow and mature and incur costs that did not scale with growth. This can lead to issues for the founders and the company as the marginal benefits received diminish in relation to the value of equity traded away. The data collected for the research in this paper does not include equity distribution for startups prior to appearing on the show as that data was not available but from a general viewership, most of the start-ups are seeking their first round of major funding. For that purpose, SharkTank is a potential area of research as all aspects of the 'deal' will have profound effects on the startup moving forward. As more financial data is made available in regards to success of startups that received funding on SharkTank, analysis can be provided to see how the deal structure impacts startups later on in their business life cycle.

In this paper, we focus on deal prediction as a first step-forward into adding knowledge to the ever-growing understanding of SharkTank and in extension, startups in general. We detail related works, deliver a new dataset - the SharkTank Deal Dataset, propose a new computational model to predict

978-1-7281-3003-3/19/$31.00 ©2019 IEEE

[2]https://www.qvc.com/lori-greiner/_/N-1z141p9/c.html

the shark-startup deal, and evaluate the results. Our later experiments on the collected STDD dataset show the superiority of our model over the baselines.

II. Related Work

Artificial Intelligence is the latest buzzword in the business world right now, with good reason. Facebook, Google, and other large tech businesses hold a treasure trove of data that had been impenetrable for quite some time until recent advancements in data science [13]. As such, the value of data has skyrocketed due to the potential use in marketing and management (as well as pretty much every part of business), leading to the acquisition of certain companies, such as LinkedIn, mainly for their data [7]. Hence the prediction that data will soon be the most valuable resource in the world [5].

But academic research has not been left behind and in a recent article, Liu *et al.* [12] applied predictive machine-learning algorithms to social media data from sources such as LinkedIn, Facebook, and Twitter to predict future career path moves. Their models had varying success between career path types - some were easier to predict than others and the results have the potential to help prepare businesses for alternative career path moves amongst their current employees. In another article, D. Singh *et al.* [3] applied different machine learning models such as support vector regression, gaussian process regression, and a multi-layer perceptron in comparison to linear regression to predict two things from two different data sets: airline passenger numbers and wine sales. The results showed that the machine learning models either had relatively similar predictions or outperformed the linear regression model. In [21], S. Kotsiantis *et al.* created a hybrid forecasting system that utilized multiple classifiers such as decision trees and artificial neural networks that were trained on different types of data (i.e. financial ratios that were broken into different categories) to forecast fraudulent financial statements. The results were very accurate with accuracies ranging from 91.2 to 93.9 percent. There are many other examples of published articles where machine learning is applied to business data and it is easy to envision how machine learning will continue to drive major developments or events in both the business and academic research world.

With all of this focus of applying AI concepts to business, it was only time until someone would apply these algorithms to startup data. When this paper was first conceptualized, the experiment in mind had to do with venture capital data as certain articles [19] [25] have interviewed venture capital managers that have integrated machine learning models into their analysis, but no academic research has been made available that relates to VC and machine learning in this sense. But a couple of issues arose, including many that are detailed in [10] in which Kaplan and Lerner compile available data and research on venture capital investments and performance to find that there are many biases that currently cannot be accounted for in existing data such as under reporting failed investments.

There is also the issue of quantifying success or failure due to the large amount of ways that businesses can take form over their lifetime. In [14], McGowan gives a definition of a successful startup which is surprisingly complex that further reinforces this idea. Similarly, Levin-Epstein [11] and De-Mers [4], introduce several signs for a successful startup, i.e., being well-funded and providing a great service or product, as well as factors such as no market need, running out of cash, strong competition, pricing/cost issues, poor product, poor marketing, ignoring customers, product miss-timing, etc., that could be possibly captured but needs great attention to detail that would not be able to be captured by current computational methods. The final nail in the coffin was a lack of publicly available dataset(s).

Instead, we decided to apply these concepts to the TV show SharkTank, which had publicly available data and a concept similar in nature to venture capital. Plenty of analysis has been published online such as articles like [16], in which Miller attempts to find if there was any gender bias in how the sharks decided to invest or the article [1], in which Breslouer attempts to dissect the art of valuation on SharkTank as well as the article [6] which provides statistics in regards to the 'pitch'. In [9], Giang goes into detail about the 'Shark Tank Effect', which is the positive impact that startups or founders receive after they appear on the show even if they did not strike a deal. There are plenty more articles that can be found online in the form of published articles or online forum posts.

To the best of our knowledge, predictive algorithms had yet to be applied to SharkTank data in the way that we proposed. In [20], Raghvendra *et al.* applied a CNN-LSTM hybrid to audio from the show to predict whether or not a start-up would be funded to good results. While the method of prediction and type of data differed than what we applied in this paper, this is a good example of how machine learning could potentially be applied to something other than traditional business data that would still impact the bottom line (in this case, finding some optimal way to interact with investors in hopes of securing a business deal).

From what has been said so far in the related works section, we can see that a standard dataset for deal prediction is missing and that there is a legitimate need to propose a computational model. Therefore, in this paper, we construct a standard dataset, named SharkTank Deal Dataset (STDD), from multiple datasets as well as information derived from episodes of SharkTank and propose a new computational model to tackle this problem.

III. Dataset and Computational Model

In the following section, more information about the data will be explained, but the idea of what we tried to accomplish was to predict whether a startup will make a deal with the 'sharks'. As such, the nature of each data record had to do with the team, product, sharks, etc. and was classified as a deal (1) or no deal (0). From this data and the results, we will further analyze in section 4 whether there are any defining patterns found within the data in relation to the prediction

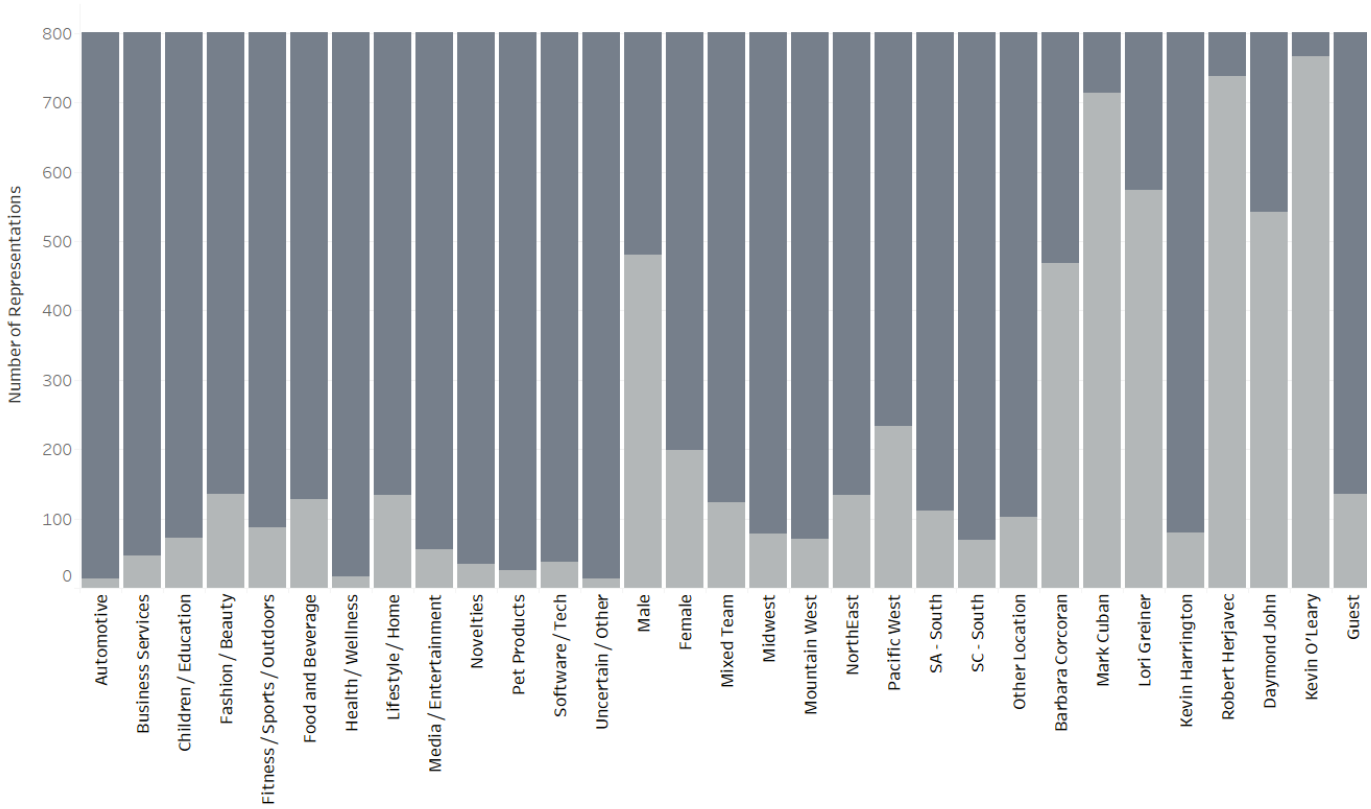

Fig. 1. The distribution of each feature dimension in our dataset.

models, the prediction accuracy rates, and the weights trained during the training phase.

A. Data Collection

The first goal of the work was to create a model and apply it to data to predict whether a specific startup was to receive a 'deal'. The emphasis of the data related towards things that were in control by the startup before appearing on the show and not about things that happened during the show such as the bargaining phase with the 'sharks' and quality of presentation.

The data collected encompassed the product category, team demographic, state origin, total amount of equity offered, the total amount asked for, and the valuation of each company. We also listed each judge that appeared on that episode, placing guest judges in 'other judges'. Lori Greiner, Mark Cuban, Barbara Corcoran, Kevin Harrington, Kevin O'Leary, Robert Herjavec, and Daymond John were given their own classification considering that they were or still are the main 'sharks' on the TV show. The data we collected came from two primary sources that we found online [22] [23]. We combined them and filled in any missing information independently through various sources such as SharkTankBlog, and transformed the data so that it could be used in our machine learning models. For product category, team demographic, judge, and geographic origin, we did binary classification (1/0). Geographic origin was decided by the state origin of each company which was then further grouped by census territory - an example would be Oregon, Washington, and California bundled into the Pacific West category. For equity offered, valuation, and amount offered, z-score values for each population were calculated for the dataset. The following Table 1 and Figure 1 are breakdowns of total categorical representations within the data set which comprised of 802 data records:

TABLE I
THE STATISTICS OF STDD DATASET.

	Size	%
Full Dataset	802	100%
Training Set	642	80%
Testing Set	160	20%
Deal	438	55%
No Deal	364	45%
Seasons	9	-

Out of 802 products that have appeared on the show, 55% struck a deal with the sharks. This number may not be entirely representative of the actual teams that received a deal considering that some start-ups may not have agreed with the

terms or some other reason such as wanting to appear on the show for marketing reasons.

A few observations from Figure 1 include the following: the 'Judge' features are consistently the most represented considering they are not mutually exclusive and there are at least four or five per show, the most represented geographic location is the Pacific West which can be explained by the inclusion of the state of California from which a significant amount of start-ups originated that appeared on the show, the majority of teams are all male which does not factor the males that were on male/female teams, and the greatest product categories represented on the show are Lifestyle/Home, Food and Beverage, and Fashion/Beauty.

For the evaluation, the data was once randomly partitioned into two sets at a ratio of 4:1 for training and testing, respectively. The dataset will be released along with the publication of this paper.

B. Computational Model

1) Problem Formulation: Similar to other prediction problems [12], [17], [18], the status of a startup at the deal/no deal decision can be linearly predicted from the input features x as follows:

$$f(x) = \omega^T x, \tag{1}$$

where ω denotes the weight for the linear mapping function of the input feature x. Given the features and the groundtruth labels mentioned in the Data set Collection section, we aim to minimize the error of the prediction and the groundtruth:

$$\min \|\omega^T X - Y\|^2, \tag{2}$$

where X are the extracted features of all startups in the training set, Y is the corresponding groundtruth labels. In addition, in order to avoid the overfitting, we apply the ℓ_2 regularization term. For the sake of the later visualization, we set a constraint of non-negative weights ω. Therefore, we define our objective function as below:

$$\min \|\omega^T X - Y\|^2 + \lambda\|\omega\|^2,\ s.t.\ \omega \geq 0. \tag{3}$$

This objective function attempts to minimize the difference between the predicted results and the groundtruth labels. Note that the regularization term with λ is added. For the sake of the later visualization, we set a constraint of non-negative weights ω.

2) Optimization: To solve Eqn. 3, we first construct $C = [X\ \lambda I]$ and $d = [Y\ 0]$. Then, we solve the non-negative least square problem $\|\omega C - d\|, s.t.\ \omega \geq 0$. We use the off-the-shelf toolbox, *i.e.*, lsqnonneg in MATLAB to efficiently solve the equation. Note that the regularization parameter λ will be decided via the grid search with different values from 0.001 to 1. For the testing phase, the trained weight vector ω will be used to predict the label y_t of the input features x_t by using Eqn. 1.

IV. Evaluation

A. Evaluation Settings

We consider the following baselines:

1) Feed-Forward Neural Network: we apply a Feed-Forward neural network with one hidden layer consisting of five nodes and trained the model using scaled conjugate gradient backpropagation. We adopt the small network to avoid the overfitting. Results are averaged over ten instances.
2) Ridge Regression: we apply Ridge Regression with regularization term similar to ours.
3) Non-Linear Least Squares: we adopt Non-Linear Least Squares regression in the same form as the second model.
4) Support Vector Regression: we utilize the existing libSVM toolbox [2].
5) K-Nearest Neighbor: we apply a K-Nearest Neighbor model trained using the FITCKNN function that can be found in the MATLAB Toolbox with different Ks ranging from 1 to 10.

For the evaluation, we use the collected STDD dataset. We use accuracy as the main evaluation metric. The accuracy rate is calculated by comparing all of the predicted values to the ground-truth values, summing the total amount of true positives and true negatives, and dividing the sum by the total amount of test records.

B. Evaluation Results

We run our model and the baselines on STDD dataset. The model accuracies are shown in Table 2 below. Our model achieves the highest prediction accuracy with 62.5%, followed by Non-Linear Least Squares at 61.25%, K-Nearest Neighbor (K = 6) at 60.62%, Support Vector Regression at 58.75%, Ridge Regression at 56.25%, K-Nearest Neighbor (K = 1) at 56.25%, and Neural Network (5 Nodes) at 55.12%. The combination of non-negative least squares and ridge regressions actually boosts the performance as seen in our model. Meanwhile, the neural network achieves the lowest accuracy. The KNN performs better than neural network with K = 6 performing the best from K = 1 to 10. The reason may be the small size of the dataset.

TABLE II
COMPARISON OF OUR PROPOSED FRAMEWORK WITH BASELINES.

Model	**Accuracy**
Neural Network (5 Nodes)	55.12%
Non-Linear Least Squares	61.25%
Ridge Regression	56.25%
Support Vector Regression	58.75%
K-Nearest Neighbor (K=1)	56.25%
K-Nearest Neighbor (K=6)	60.62%
Our Model	**62.50%**

In theory, the results of the model would have the potential of helping startups figure out their chances of striking a

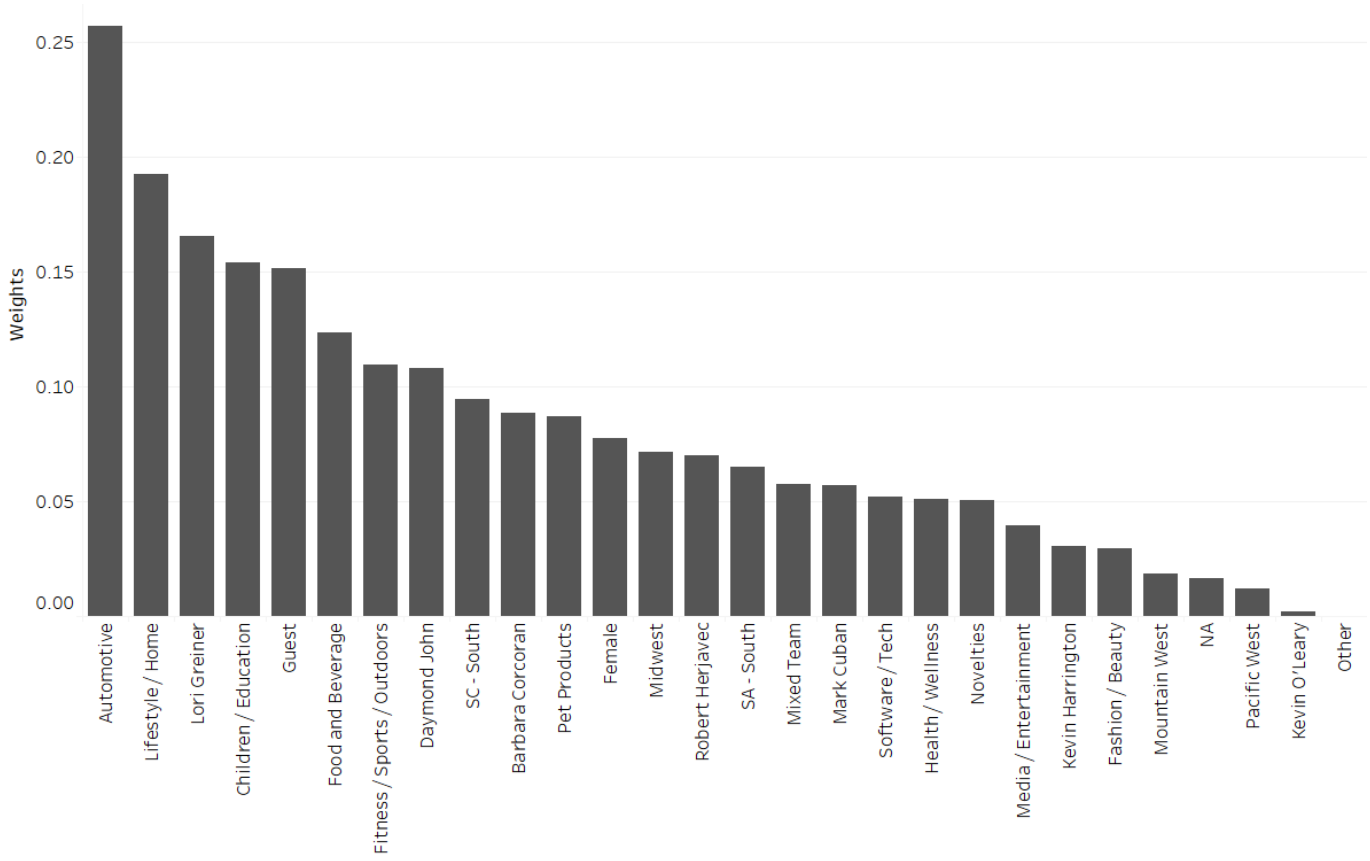

Fig. 2. Learned feature weights from our model.

deal before applying for the show or deciding the optimal valuation and amount of equity to offer. At a closer look, we visualize the learned weights ω from our model. As shown in Figure 2, features of certain categories and judges such as Automotive, Lori Greiner, and Lifestyle/Home play an important role into the prediction result of 'deal' while features such as judge Mark Cuban, demographic 'Male', and geographic area 'Northeast' have no impact.

We further calculate the correlation coefficients for all features and plotted them against the weights in Figure 2. From these results, we can see that the correlation coefficients closely related to features with the largest weights are positive and have the highest values whereas correlation coefficients related to features with weights that are closer to the middle have greater variability especially in regards to sign. And finally, correlation coefficients related to features with the smallest weights are negative and have the largest negative values of all the features.

V. Conclusions and Future Work

In this paper, we introduced a new STDD dataset for startup deal prediction. We also proposed a new model with a non-negative weight constraint and a regularization term. Our proposed model performed well and achieved the best performance on the dataset out of all of the models. We analyzed the results as well as the weights attached to each feature and concluded that they are in line with general viewership.

In the future, we aim to extend the dataset with future seasons of SharkTank as well as seasons from international versions and apply different deep learning models for the task. There is still room for the improvement in the computational framework. In addition, our dataset can be extended with more information, for example, if a startup received an offer from a judge but did not accept, if a startup received an offer that was better or worse than the initial offer, how many judges offered a deal, and descriptive factors of presentations such as number of presenters and length of presentation amongst others. There is also the possibility of providing different forms of analysis beyond classification such as predicting the difference between what was dealt versus what was initially asked for. Not only would this information provide greater clarity as to the end result, but could also explain how important the presentation really is and what factors of the presentation are important in order to potentially increase the accuracy of our models.

Acknowledgement

We are grateful to the NVIDIA corporation for supporting our research in this area.

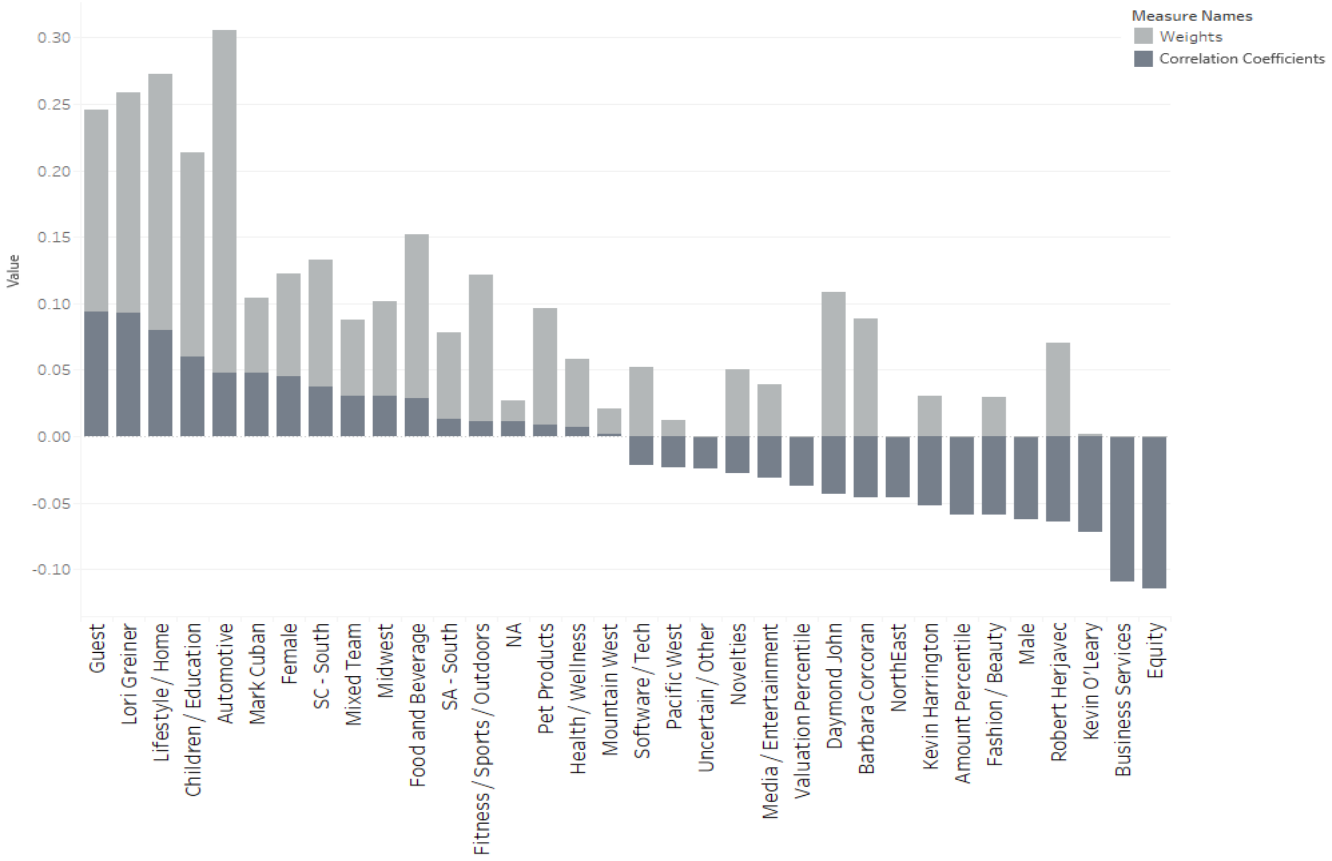

Fig. 3. Feature Correlation Coefficients and Feature Weights.

REFERENCES

[1] L. Breslouer. This is why the sharks get so hung up on valuation during 'shark tank'. *Thrillist*, page 10, 2017.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[3] S. M. R. Y. D. Asir Antony Gnana Singh, E. Jebamalar Leavline. Machine learning based business forecasting. *I.J. Information Engineering and Electronic Business*, 6:40–51, 2018.

[4] J. DeMers. A new study reveals the 20 factors that predict startup failure: Do any apply to you? *Entrepreneur*, page 5, 2018.

[5] T. Economist. The worlds most valuable resource is no longer oil, but data. *The Economist*, page 4, 2017.

[6] D. Exits. How to pitch a shark: The main factors that lead to a deal on shark tank. *Digital Exits*, page 5, 18.

[7] M. J. Foley. Why microsoft just bought linkedin: It's all about the data. *Zdnet.com*, page 3, 2016.

[8] D. Gage. The venture capital secret: 3 out of 4 start-ups fail. *The Wall Street Journal*, 2012.

[9] V. Giang. The "shark tank" effect: How even the show's losers became winners. *American Express*, 2014.

[10] S. N. Kaplan and J. Lerner. Venture capital data: Opportunities and challenges. Working Paper 22500, National Bureau of Economic Research, August 2016.

[11] A. Levin-Epstein. Joining a startup? 6 signs it'll be a success. *CBS*, page 3, 2013.

[12] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum. Fortune teller: Predicting your career path. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 201–207, 2016.

[13] B. Marr. A short history of machine learning – every manager should read. *Forbes*, page 4, 2016.

[14] E. McGowan. The definition of a successful startup is surprisingly complex. *Startups.co*, page 19, 2018.

[15] R. Merlino. Sharktankblog, 2019.

[16] J. Miller. Swimming or sinking in the shark tank...does gender matter? *Medium*, page 7, 2018.

[17] T. V. Nguyen, S. Liu, B. Ni, J. Tan, Y. Rui, and S. Yan. Sense beauty via face, dressing, and/or voice. In *Proceedings of ACM Multimedia Conference*, pages 239–248, 2012.

[18] T. V. Nguyen, S. Liu, B. Ni, J. Tan, Y. Rui, and S. Yan. Towards decrypting attractiveness via multi-modality cues. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 9(4):28:1–28:20, 2013.

[19] T. Quarterly. A machine-learning approach to venture capital. *McKinsey Quarterly*, page 10, 2017.

[20] S. Raghvendra, J. Wood, and M. Ziao. Pitch perfect: Predicting startup funding success based on shark tank audio. page 8, 2017.

[21] D. T. S. Kotsiantis, E. Koumanakos and V. Tampakas. Forecasting fraudulent financial statements using data mining. *International Journal of Computational Intelligence*, 3, 2006.

[22] R. Sathyajit. Shark tank pitches. *Kaggle*, page 1, 2017.

[23] H. Tecco. What have you learned from watching the television program shark tank? *Quora*, page 2, 2015.

[24] R. Varshneya. How shark tank is revolutionizing business school. *Entrepreneur.com*, page 4, 2017.

[25] E. Vidra. How venture capital funds leverage ai and big data. *VC Cafe*, page 7, 2018.

[26] J. Xavier. 75 percent of startups fail, but it's no biggie. *BizJournals*, page 2, 2012.

[27] B. Zider. How venture capital works. *Harvard Business Review*, 1998.