DECODING SUCCESS ON SHARK TANK

Analyzing Key Predictors of Funding Success

Shuxi Chen

Fall 2024 - MGSC-661 | 2024.12.04

1. Introduction

Securing funding is a critical milestone for startups, as access to capital often determines their ability to grow and succeed. However, the process of convincing investors or venture capitalists, such as those on Shark Tank, is highly competitive and requires founders to present their business in the most compelling manner. Understanding the key factors that influence funding success has become a significant area of interest for entrepreneurs seeking to refine their strategies.

The pitch serves as a pivotal element in a startup's journey toward funding. It is the primary medium through which entrepreneurs convey their vision, demonstrate potential, and appeal to investors' interests. The success of a pitch is influenced by multiple factors, including the nature of the product or service offered, its market potential, the clarity and creativity of the pitch itself, the financial strategy presented, and the preferences of the investors. Timing and contextual factors, such as location or industry trends, also play a significant role in determining outcomes.

This project aims to predict whether a startup will secure a deal with the "sharks" on Shark Tank by leveraging historical data and advanced classification techniques. Through clustering, dimensionality reduction, and text analysis, the study seeks to uncover patterns and identify the most significant variables that contribute to successful outcomes. These findings not only offer a deeper understanding of the funding process but also provide actionable insights for both entrepreneurs and investors. By highlighting what matters most in the pitching process, the analysis aspires to optimize decision-making and improve the likelihood of success for startups navigating this competitive landscape.

2. Data Description

The dataset is a total of 19 variables across each episode of the first 6 seasons with 495 pitches (entrepreneur entities), to approach them in a logical way, they can be simplified into 6 categories as following data profile:

Table A: Detailed description of dataset collected from Kaggle

Category	Variable Name	Type	Description
Business	title	character	The title/name of the pitched business/product.
	description	character	A brief description of the business/product pitched.
	category	character	The industry or niche of the business.
	location	character	The location of the business/entrepreneurs.
	website	character	The business's website (if available).
Outcome	deal	logical	Indicates whether a deal was made (True/False).
Entrepreneur	entrepreneurs	character	Names of the entrepreneur(s) pitching the idea.
	Multiple Entrepreneurs	logical	Indicates if there were multiple entrepreneurs (True/False).

Financial	askedFor	Numerical	The amount of money requested by the entrepreneur.
	exchangeForStake	Numerical	The equity stake offered in exchange for the investment.
	valuation	Numerical	The implied valuation based on the ask and offered equity.
Show Metadata	episode	character	The episode number.
	season	character	The season number.
	episode-season	character	A combination of season and episode numbers.
Shark	shark1	character	Names of the sharks present during the pitch.
	shark2	character	Names of the sharks present during the pitch.
	shark3	character	Names of the sharks present during the pitch.
	shark4	character	Names of the sharks present during the pitch.
	shark5	character	Names of the sharks present during the pitch.

Before initiating the analysis, addressing data quality issues was essential. Upon examination, no duplicate entries were identified in the dataset. However, missing values were observed in the variables **website** and **entrepreneurs**. These were excluded from further analysis as they did not contribute value to the predictive model in this specific context, despite their potential relevance in other settings. The subsequent analysis is organized according to the categories defined in **Table A**, ensuring a systematic exploration of the dataset.

2.1 Descriptive Analysis

- Outcome

The target variable, deal, represents whether a participant secured a deal or not. A balanced distribution was observed, with 50.7% of participants securing a deal and 49.3% not securing a deal. This balance indicates a healthy mix of successful and unsuccessful pitches.

- Business Information - Location

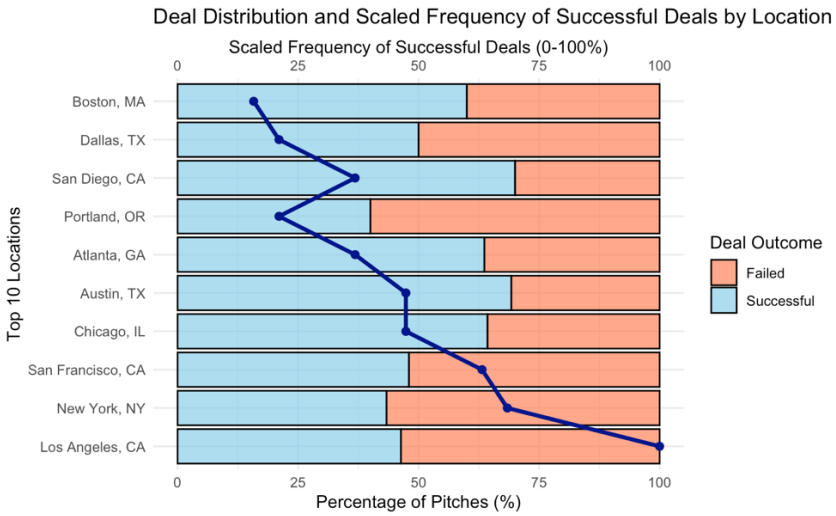


Figure 1: Deal Distribution and Scaled Frequency of Successful Deals by Top 10 Locations

Los Angeles has the highest relative frequency of successful deals, though its success rate is slightly below 50%. In contrast, cities like New York and San Francisco have lower proportions of successful deals compared to their total pitches.

Given the 51 unique levels in the location feature, analyzing success at the state or city level can be fragmented and less informative. To address this, geographic origins were grouped by Census Division as states within the same divisions often share similar economic structures, industry focus, and market conditions.

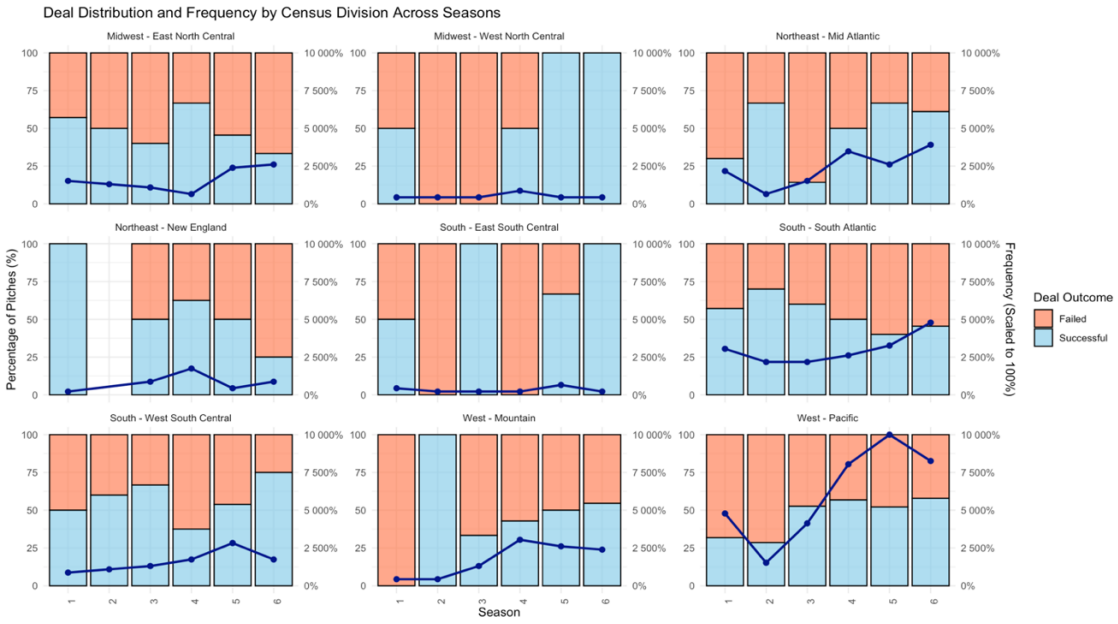


Figure 2: Percentage of successful and failed deals for each Census Division across seasons

The Pacific West, led by California, is the most represented region, followed by the South Atlantic and Mid Atlantic divisions. In terms of success rate trends, the Pacific West and South Atlantic show relatively balanced success and failure rates, whereas the Mid Atlantic and West South Central regions exhibit greater fluctuations in success rates across seasons.

These noticeable variations highlight the effectiveness of this new grouping rule, and potential interaction effects between **location** and **season**.

- Business Information – Category

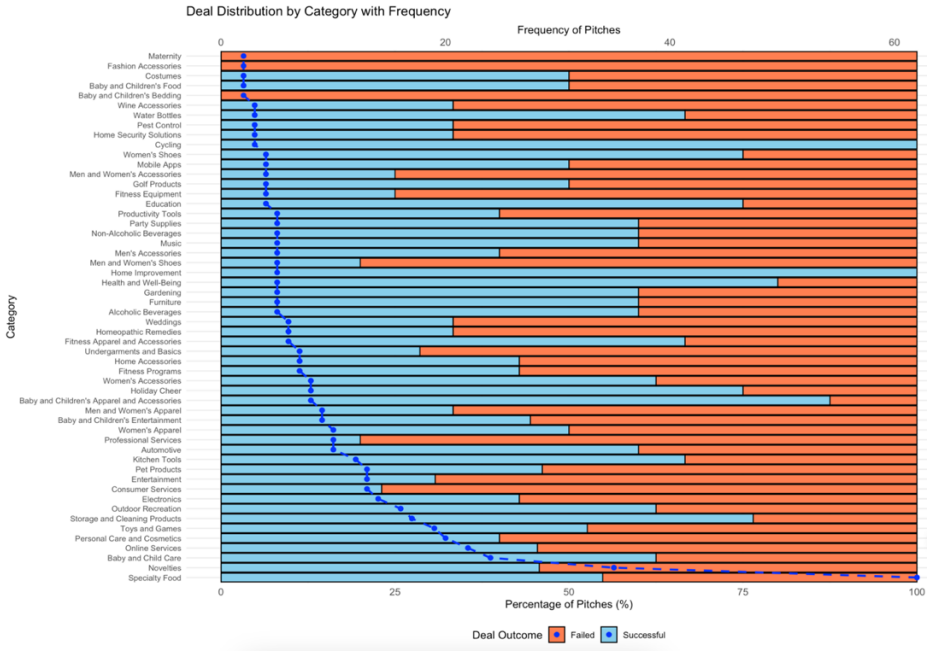


Figure 3: Deal Distribution by Category with frequency

Specialty Food emerges as the leading category, followed by Novelties and other significant sectors such as Baby and Children's Care and Personal Care and Cosmetics. High-performing categories include Cycling, Home Improvement, and Baby and Children's Apparel and Accessories, which consistently achieve the highest success rates.

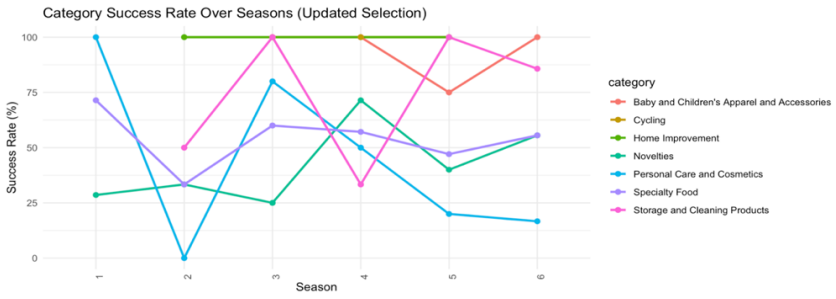


Figure 4: Selecting Categories Performance Over Time

Success rates vary across categories, with some demonstrating consistent appeal to investors while others face challenges. Categories such as Home Improvement and Cycling show strong performance and sustained investor appeal over time, whereas Novelties and Personal Care and Cosmetics often struggle to secure investments.

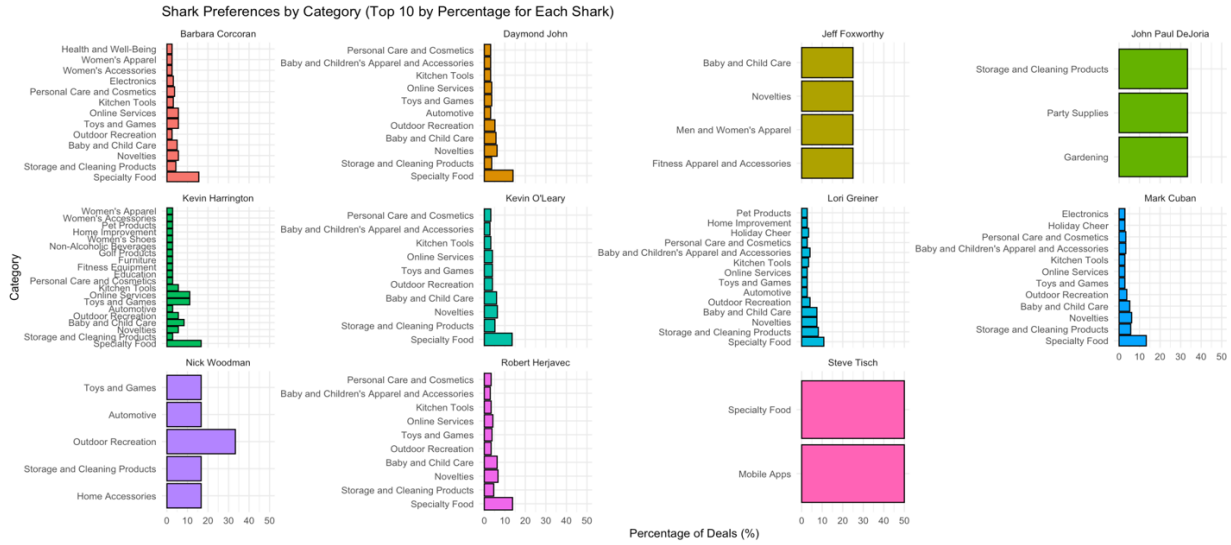


Figure 5: Top 10 categories for each shark based on the percentage of deals

An analysis of the top 10 categories per shark reveals that, except for guest sharks, most sharks' investment portfolios follow similar trends. Specialty Food is the most popular category, reflecting high deal frequency and broad appeal across all sharks.

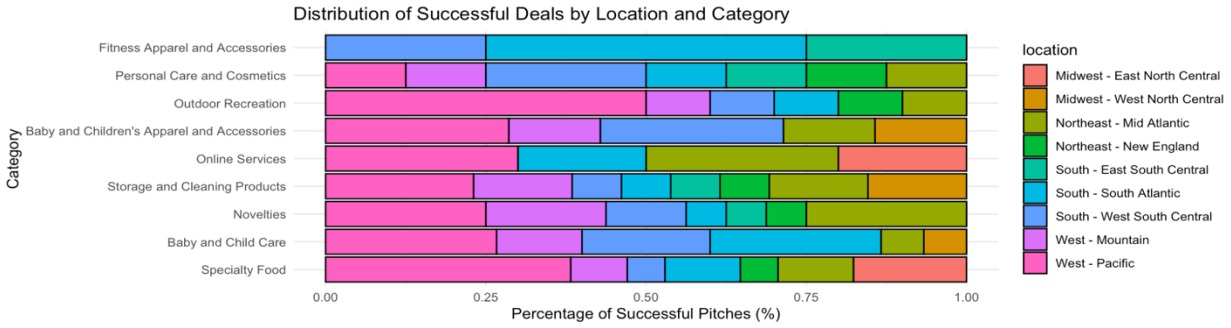


Figure 6: Category Performance by Location

The success of business categories also varies by region, with the West Pacific leading overall, followed by the South Atlantic and Mid Atlantic regions. Certain categories, such as Specialty Food, Online Services, and Storage and Cleaning Products, exhibit strong regional preferences.

From above, the **main categories** are identified: Specialty Food, Storage and Cleaning Products, Baby and Child Care, Novelties, Baby and Children's Apparel and Accessories, Personal Care and Cosmetics, Outdoor Recreation, Online Services, Fitness Apparel and Accessories.

These trends suggest potential interaction effects between **category and season** as well as **category and location**, both of which may play significant roles in predicting deal outcomes. However, no significant interaction effect is observed between **category and shark.**

- Entrepreneur Information

The variable Multiple Entrepreneurs was derived from the entrepreneurs column. Initially, it was marked as True if the field contained the separator "and." However, upon closer inspection, additional separators such as "," and "&" were identified as indicators of multiple entrepreneurs. The dataset was updated accordingly to ensure accuracy.

Regarding the entrepreneurs' names, only 8 individuals were found to have pitched twice, with the remaining founders making their first appearances. This suggests that familiarity with Shark Tank pitching is unlikely to significantly influence the outcome. Furthermore, prior research highlights the existence of gender, age, and racial biases in investor decisions. While entrepreneur names could serve as a proxy for demographic traits there is no theoretical basis to expect names to play a role. To avoid introducing unintended biases, the entrepreneurs variable was excluded from the analysis.

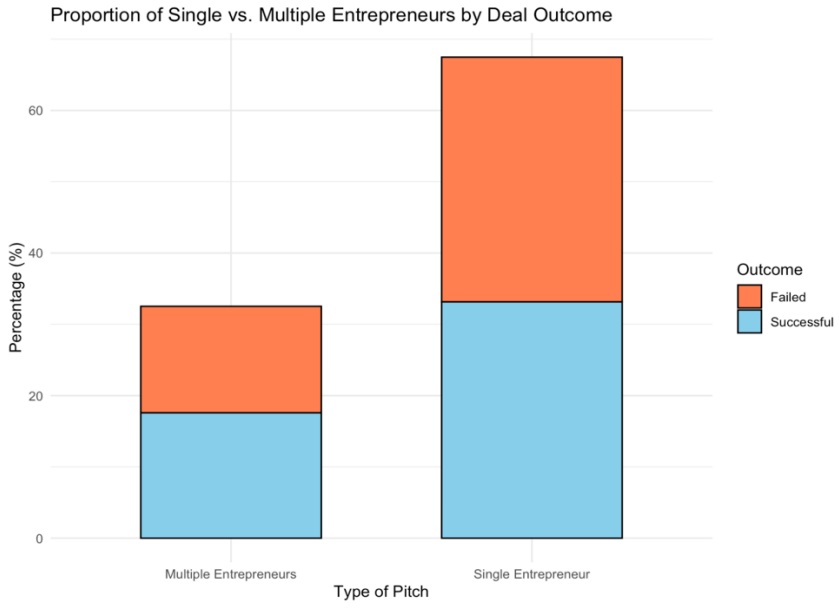


Figure 7: Proportion of Single vs. Multiple Entrepreneurs by Deal Outcome

The analysis revealed that the majority of pitches (67.5%) were made by single entrepreneurs, while 32.5% involved multiple entrepreneurs. However, the nearly 50-50 split in outcomes suggests that the type of pitch may **not** serve as a strong predictor for deal outcomes.

- Show Metadata

To analyze success rates over time, we examined the **success rate** (successful pitches compared to failed pitches) and the **scaled frequency of successful deals** (frequency of successful deals normalized relative to their maximum value), enabling fair comparisons across seasons irrespective of total pitch counts.

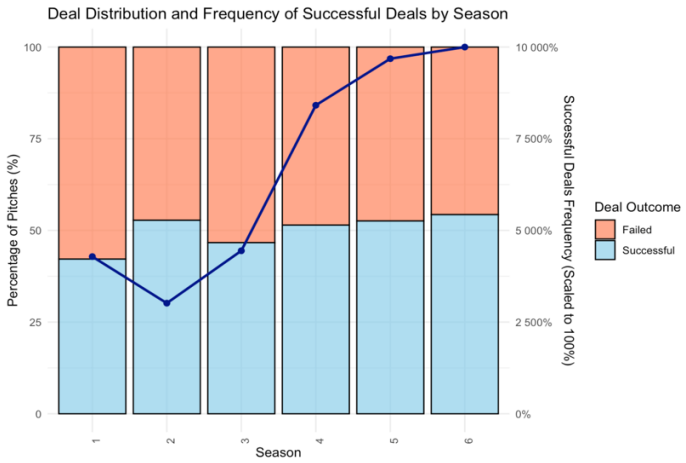


Figure 8: Deal Distribution and Frequency of Successful Deals by Season

Overall, success and failure rates appear relatively balanced across 4-6 seasons, but fluctuated in first 3 seasons. Notably, Season 2, despite having fewer total pitches, demonstrates a high success percentage. Starting from Season 3, there is a noticeable increase in the frequency of successful deals. This may suggest the **season** variable could serve as a potential predictor.

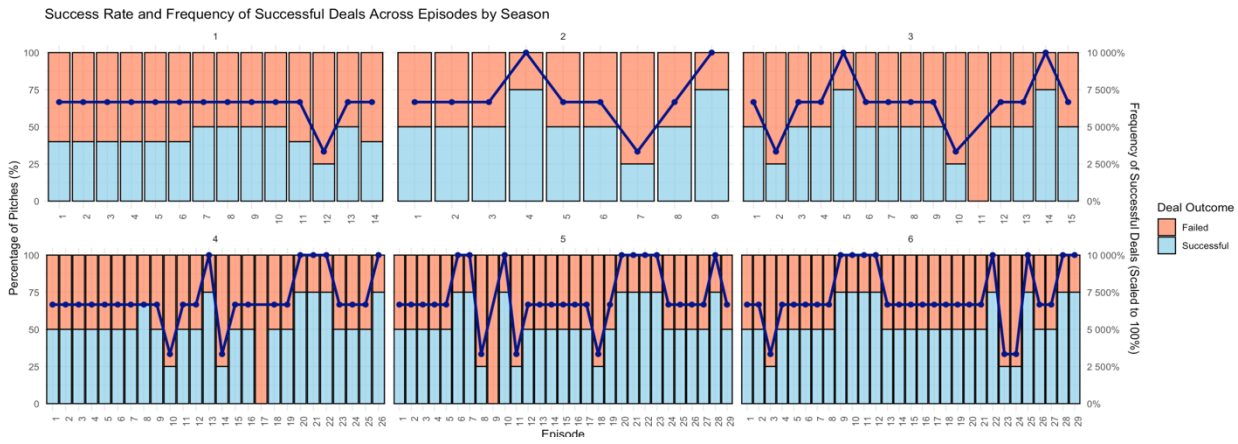


Figure 9: Success Rate and Frequency of Successful Deals Across Episodes by Season

While the scaled frequency highlights trends over time, granular analysis shows that the variation is strongly influenced by the number of episodes per season, with later seasons often exceeding 25 episodes compared to Season 2's 9 episodes. Most seasons exhibit two peaks in successful deal frequency—one in the first half and another in the second—except for Season 1, which appears more conservative, likely due to the show establishing itself as public entertainment.

This analysis suggests that **episode count** may not be a reliable predictor of success, as it is influenced by structural differences between seasons. Similarly, the episode-season feature, being derived from episode and season, is unlikely to add significant value. However, combining seasons into two broader categories—**early seasons** (1–3) and **later seasons** (4–6)—could enhance prediction models by capturing differences in show dynamics.

- Shark Information

By analyzing consistent attendance across episodes, we can distinguish between **main sharks** and **guest sharks**.

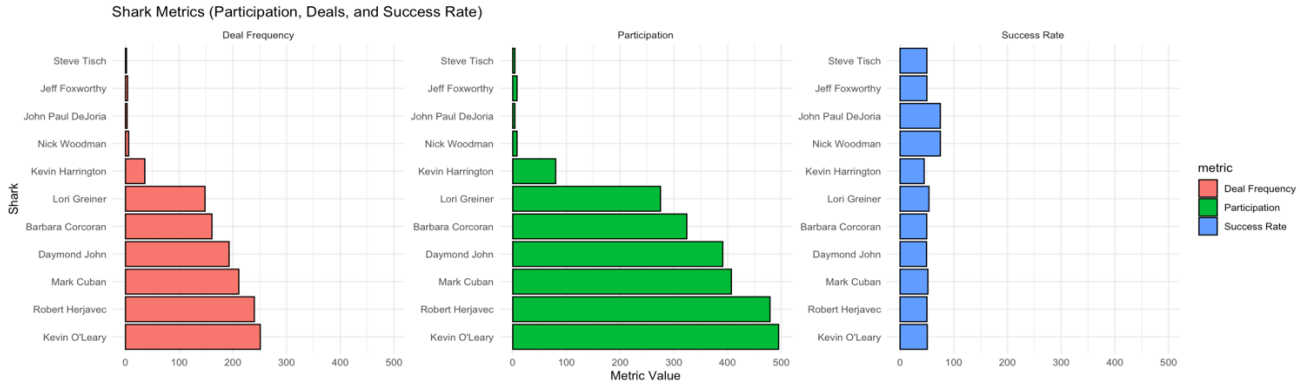


Figure 10: Distribution of sharks categorized by deal frequency, participation, success rate

The three metrics—deal frequency, participation, and success rate (calculated as the number of successful pitch episodes divided by total participation)—capture sharks' overall behaviors, including involvement, effectiveness, and investment patterns.

Main sharks (Kevin O'Leary, Robert Herjavec, Mark Cuban, Daymond John, and Barbara Corcoran) consistently participate in pitches across episodes, while **guest sharks** (John Paul DeJoria, Nick Woodman, and Steve Tisch) are often tied to specific pitches or fill in when a main shark is unavailable.

Interestingly, success rates are nearly identical across sharks, regardless of participation levels, suggesting that participation frequency may not significantly impact deal success.

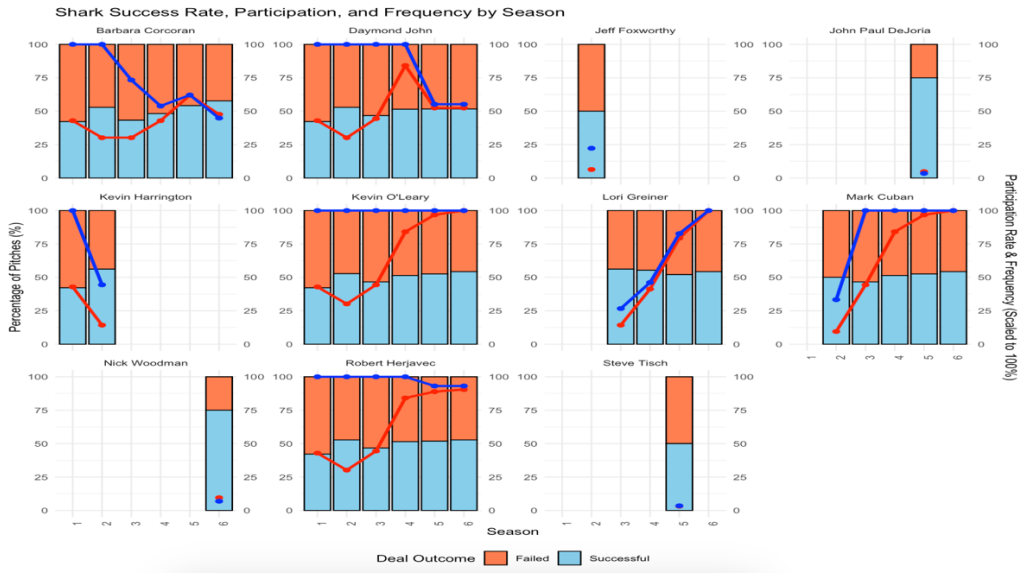


Figure 11: Distribution of sharks categorized by deal frequency, participation, success rate

Figure 11 reveals that participation and success rates vary across sharks, with main sharks consistently involved and guest sharks participating less frequently. Lori Greiner and Mark Cuban are identified as aggressive investors, while Kevin O'Leary and Robert Herjavec show increased activity in later seasons. Despite participation fluctuations, success rates remain stable, with no evidence of conservative trends. Consequently, converting shark names into predictors is reasonable, while removing constant presence sharks (e.g., Kevin O'Leary) is necessary as they lack predictive value. Plus, interaction effects between **sharks and seasons** are not required for the model.

- Financial Information

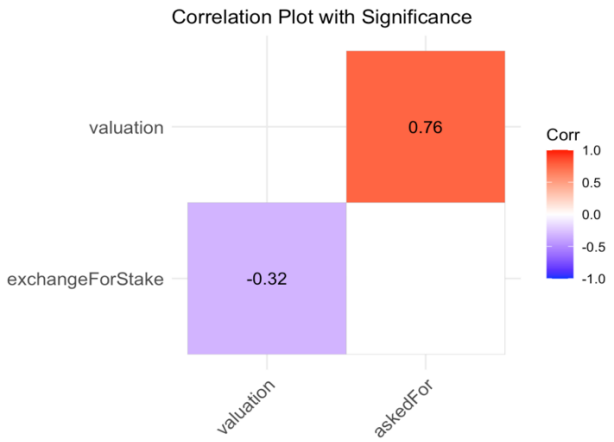


Figure 12: Correlation Plot Between Valuation, AskedFor, and ExchangeForStake

Valuation is calculated by dividing **askedFor** by (**exchangeForStake** / 100), which inherently creates a relationship between these variables. The correlation plot confirms this, showing a strong positive correlation between **valuation** and **askedFor**, and a moderate negative correlation between **valuation and exchangeForStake**.

Including all three variables introduces redundancy and collinearity. Since founders and investors can directly control askedFor and exchangeForStake, analyzing these variables provides actionable insights, and the lack of a significant correlation between them supports their independent contributions. Therefore, askedFor and exchangeForStake were retained, while valuation was excluded from the model.

2.2 PCA and Hierarchical clustering - category

Given the high dimensionality of the category variable, all categories outside the identified main ones were grouped into clusters using **hierarchical clustering** and **PCA**. Each category was quantified using engineered features such as **success rate**, **total pitches**, **distribution across regions**, **average episode presence**, **unique sharks**, **average deal size**, and **average requested equity**, enabling meaningful comparisons between categories and identifying shared characteristics for effective grouping.

Results of Hierarchical Clustering

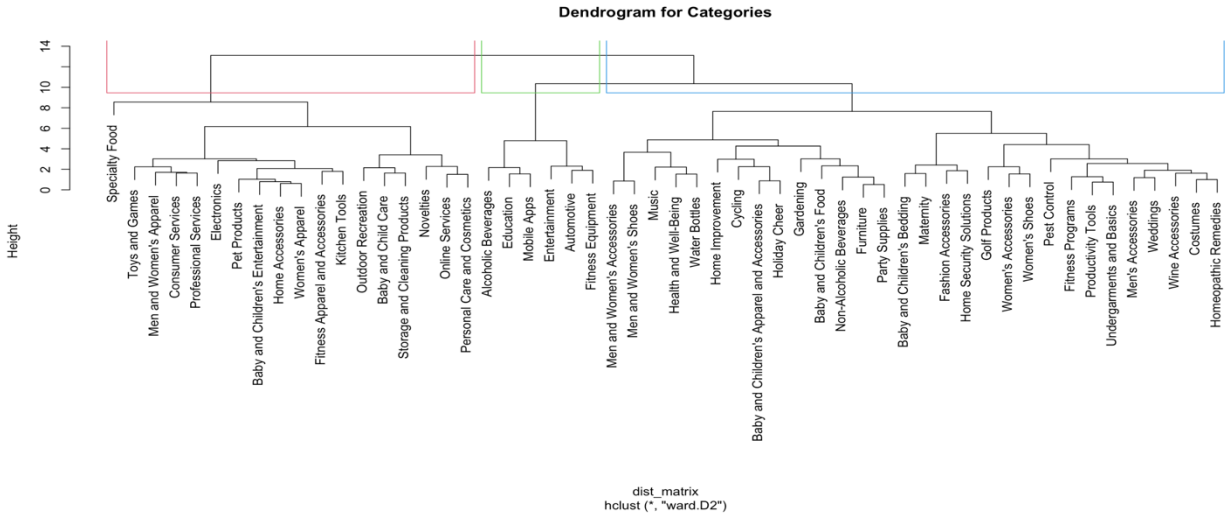


Figure 13: Dendrogram for Categories Based on Hierarchical Clustering

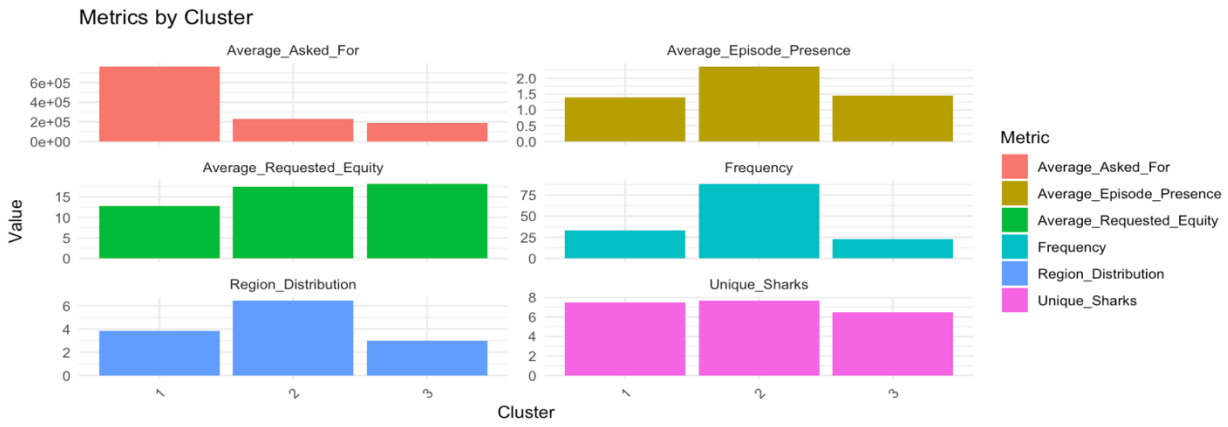


Figure 14: Clusters Characteristics by Metrics

The hierarchical clustering model identified three distinct groups:

- **Cluster 1**: High **Asked For** ($761,089.7), moderate **Requested Equity** (12.72%), and regional focus (3.83 regions). Moderate **Frequency** (33.33) and **Unique Sharks** (7.5).
- **Cluster 2**: Lower **Asked For** ($230,985.2), higher **Requested Equity** (17.51%), broadest regional spread (6.44 regions), and highest **Frequency** (88.33).
- **Cluster 3**: Lowest **Asked For** ($191,952.4), highest **Requested Equity** (18.2%), narrow regional focus (3 regions), and lowest **Frequency** (22.83).

Results of PCA

The PCA analysis suggests four groups, with the first four principal components explaining nearly 90% of the variance. For more detailed PCA results, refer to **Appendix Figure 1** and **Appendix Figure 2.**

While both methods highlighted unique characteristics between groups, the **clustering groups showed better model performance**. Therefore, the hierarchical clustering results were adopted for the final category groupings.

2.3 Text Analysis – description

The description provides a detailed narrative of the product or business idea, emphasizing features, value propositions, and unique selling points, while capturing the tone and goals of the founders' pitch. Unlike titles, which are brief and designed for attention, descriptions offer greater depth, making **word count** a key feature for evaluating pitch complexity. As such, titles were excluded to focus on analyzing descriptions for **sentiment** and **word count**.

To ensure a robust evaluation of the descriptive content, two dictionaries were employed. Specifically, one is **NRC**, which evaluates cognitive and emotional states, assessing

psychological attributes such as sociability and morality, categorized into positive and negative dimensions. The other is **Hedonometer**, which includes 10,222 commonly used words. Focuses on overall happiness or sentiment score derived from word usage.

During text preprocessing, the data was prepared through steps such as lowercasing, punctuation removal, tokenization, removal of non-alphanumeric characters, stopword removal, and lemmatization. Sentiment analysis involved counting words in predefined sentiment and psychological categories, capturing positive and negative contributions to dimensions like **Happiness**, **Morality**, and **Sociability**. These counts were normalized by dividing them by the total word count, ensuring comparability across texts. Positive and negative counts were aggregated to generate valence-related predictors, reducing dimensionality. NA values in sentiment features, arising when no words matched a category, were interpreted as no contribution from that category.

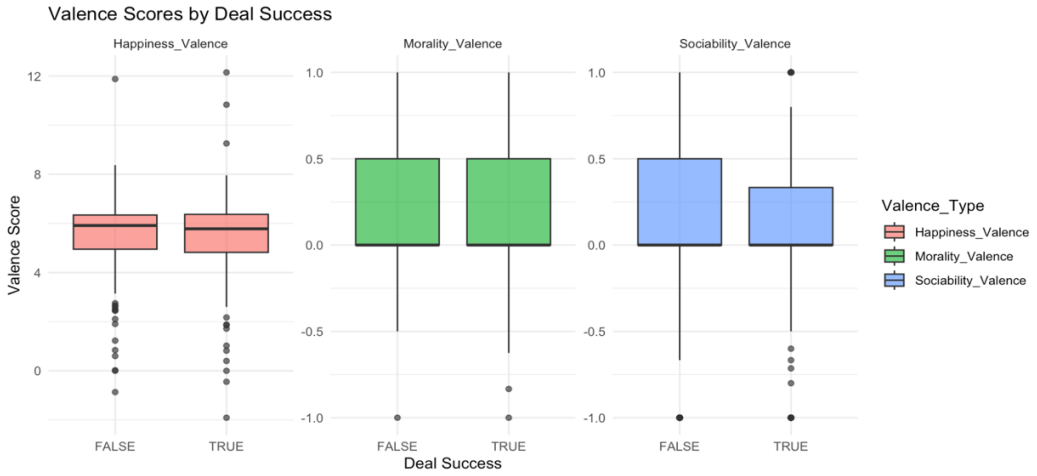


Figure 11: Valence Scores by Deal Success

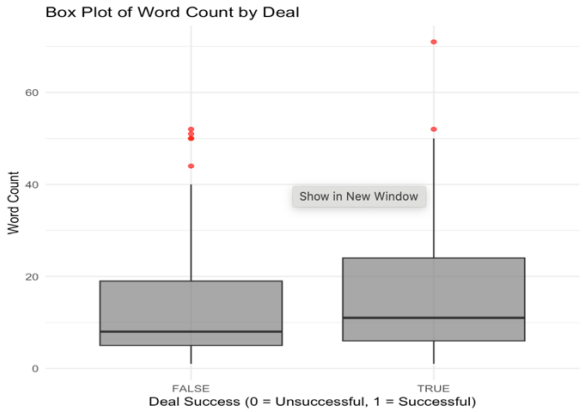


Figure 12: Box Plot of Word Count by Deal Success

The distribution shows that **Happiness Valence** exhibits the highest variance and range, while **Morality Valence** and **Sociability Valence** are more narrowly distributed. Although the

comparison of valence scores for successful and unsuccessful deals reveals no significant visual differences across these dimensions, they may still contribute predictive value when integrated with other features in the classification models.

The comparison of word count shows a slight increase in the median word count for successful pitches, indicating that this feature could have predictive value when included in the models.

2.4 Pearson's Correlation Analysis

One-hot encoding and min-max standardization were applied to prepare the data for correlation analysis.

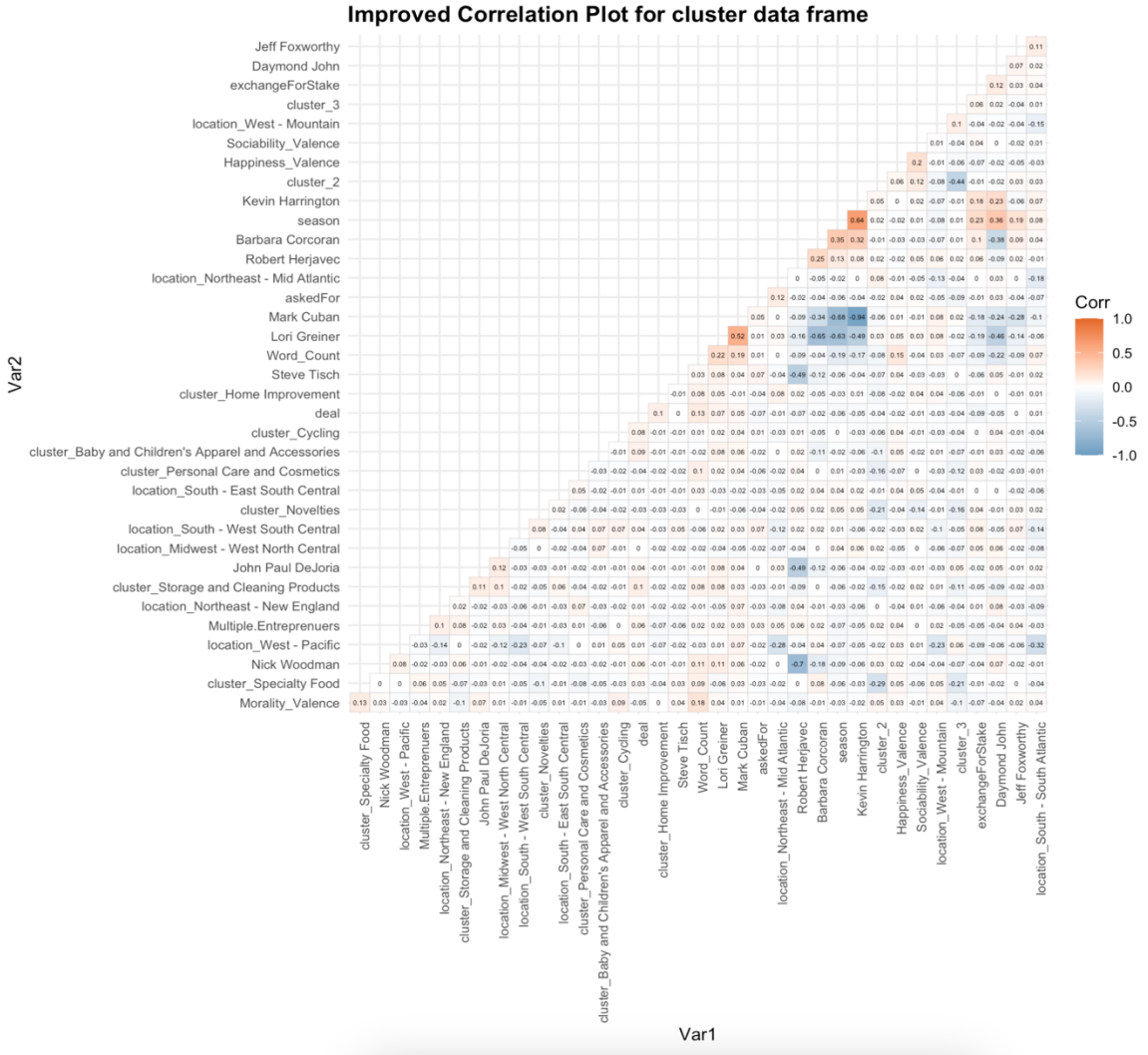


Figure 13: Correlation matrix for all variables

The analysis revealed a near-perfect negative correlation between **Kevin Harrington** and **Mark Cuban**, reflecting Mark Cuban's replacement of Kevin Harrington as a main shark. Retaining

both predictors introduces redundancy, so one was removed to resolve multicollinearity. Additionally, guest sharks were excluded due to their limited appearances.

A **VIF test** indicated remaining dependencies among some predictors. The alias() output showed that **Barbara Corcoran** is perfectly negatively correlated with both **Lori Greiner** and **Daymond John**. Lori Greiner, although a main shark, was dropped as she appeared only in later seasons, helping further reduce multicollinearity.

3. Model Selection & Methodology

To ensure reliable performance metrics, the dataset was divided into training and testing sets, with 80% allocated for training and 20% reserved for evaluation. A **stratified splitting method** was employed to preserve the original class distribution and prevent class imbalance issues.

To identify the most important predictors, **Gini importance values** were derived from a Random Forest model. This approach was chosen as it captures complex, non-linear relationships between features and the target variable, providing deeper insights into feature importance.

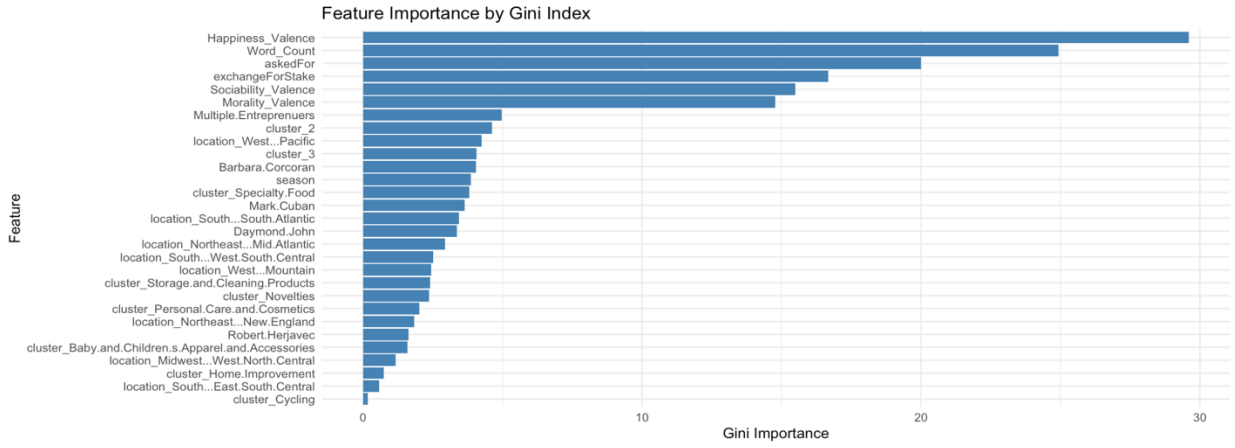


Figure 13: Feature importance derived from the Gini-values

Given the relatively small dataset, dimensionality was addressed by excluding variables with Gini scores less than 4 (top 10), forming retained variables for the model. Notably, while EDA suggested that the presence of multiple entrepreneurs was not a significant predictor, it was included in the model due to its ranking in the top 10 Gini importance scores, highlighting its potential contribution to predictive performance.

For model selection, running the models with all variables established a baseline performance metric to compare against models with selected features. **Logistic Regression** was chosen as a starting point for its simplicity and ability to provide interpretable coefficients. **Classification Trees** were used for interpretable, non-linear decision-making, while **Regression Trees** allowed for flexible modeling of continuous predictors. **Random Forest** was selected for its ability to

handle high-dimensional data and capture complex, non-linear relationships, and **Boosting algorithms like XGBoost** were included to improve overall model accuracy.

4. Results

I began with a Logistic Regression model, including all first-order variables without interaction terms, which achieved 60.5% accuracy, though only a few variables were statistically significant. When rerunning the model with selected variables, accuracy slightly dropped to 59%, confirming the effectiveness of feature selection in maintaining performance while reducing dimensionality. However, only **Word_Count** and **Morality_Valence** were significant in the model. Including identified interaction terms (location and season, category and season, category and location) improved performance by just 0.8%, suggesting that these interactions might have limited predictive power in this context. To address insignificant coefficients and enhance generalization, I applied Lasso normalization, but it did not yield any improvements.

Recognizing Logistic Regression's limitations in capturing non-linear relationships, I explored Decision Trees, Random Forest, and XGBoost models. These models were optimized using Grid Search to systematically evaluate hyperparameters, including tree depth, minimum samples for splits, number of estimators for Random Forest, and learning rate for XGBoost. To ensure consistent evaluation and mitigate overfitting during tuning, I employed 10-fold Cross-Validation. Regularization techniques, such as L1 (Lasso) and L2 (Ridge), were applied to enhance model generalization. I tested six combinations of interaction terms across these models, with the interaction between **category** and **location** consistently emerging as the most effective, while the other two interaction pairs offered only mild performance improvements. For interpretability, higher-order variables were excluded, as my primary objective was to derive actionable insights for business strategies. Therefore, the subsequent analysis focuses on first-order terms and the interaction between **category** and **location**.

Table B provides a summary of the model performances during the training stage. For a detailed summary of optimal hyperparameter settings, refer to **Appendix Table A**.

Table B: Performance metrics for different models in training stage

Model	Accuracy	Precision	Recall
Logistic regression	0.5899	0.5815	0.5573
Classification Tree	0.6253	0.6089	0.6406
Random Forest	0.9975	1.0000	0.9948
XGBoost	0.7468	0.7556	0.7083

From **Table B**, it’s evident that the Random Forest model risks overfitting, boasting an almost perfect accuracy score of nearly 1. The other models fall within a safer range, demonstrating

moderate and balanced performance. Among them, XGBoost stands out as the best-performing model, delivering promising accuracy scores. Despite my efforts to achieve reliable performance, these results appear somewhat uncertain. Therefore, I will proceed to evaluate all models on the testing stage to ensure robustness and identify the most reliable option

Table C: Performance metrics for different models in testing stage

Model	Accuracy	Precision	Recall
Logistic regression	0.5102	0.5349	0.4510
Classification Tree	0.5204	0.5357	0.5882
Random Forest	0.4796	0.5000	0.4314
XGBoost	**0.5612**	**0.5800**	**0.5686**

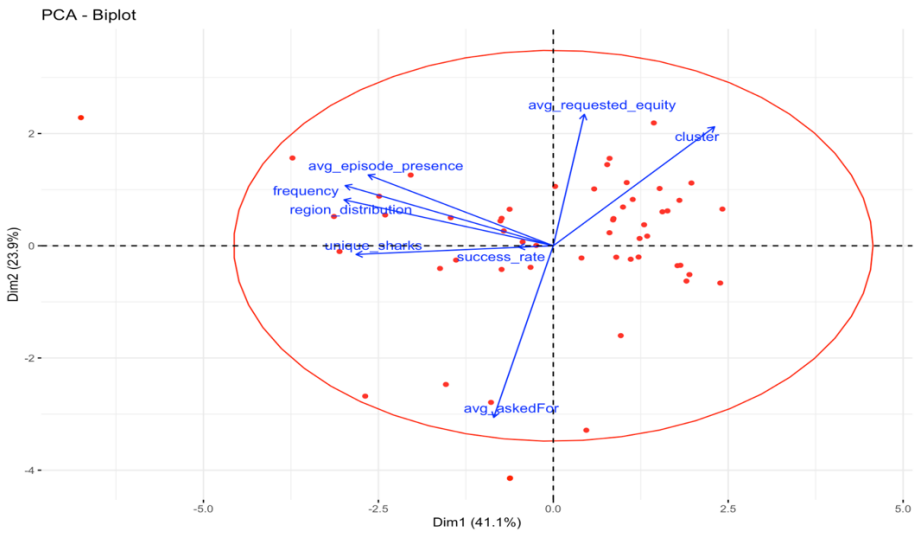
The result showed that **XGBoost** the best model, with the **final variables** being as follow: Happiness_Valence, Word Count, askedFor, exchangeForStake, Sociability_Valence, Morality_Valence, Multiple.Entreprenuers, location_West Pacific, cluster_2, cluster_3, and cluster_2* location_West Pacific, cluster_3* location_West Pacific.

5. Classification/predictions and conclusions

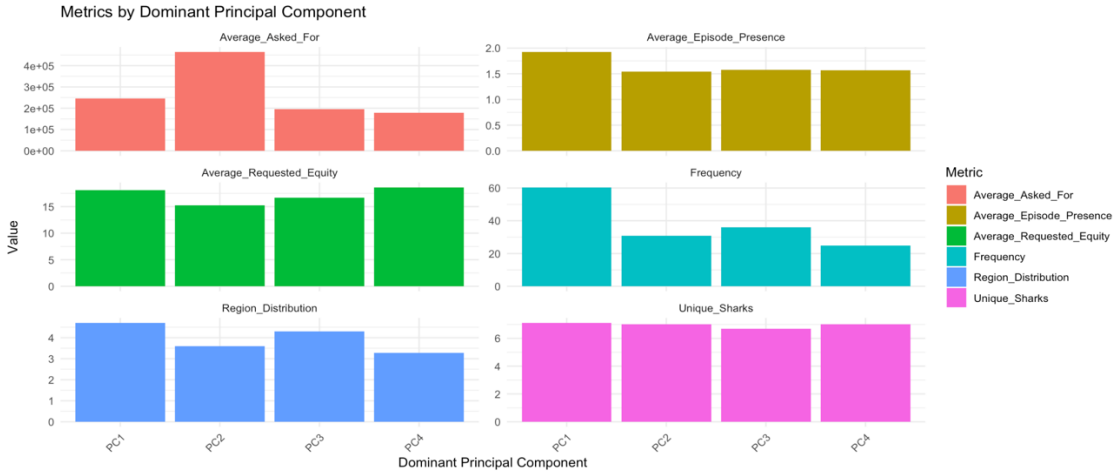
Among the models tested, XGBoost emerged as the most effective, demonstrating strong performance in both training and testing stages. It successfully balanced accuracy, precision, and recall while preserving interpretability and predictive power. The final model identified key predictors, underscoring the importance of regional and categorical dynamics in determining funding success. Entrepreneurs can enhance their chances of securing deals by crafting clear and compelling narratives, as evidenced by the significance of Word Count, and incorporating emotional and moral appeals into their pitches. For investors, recognizing high-performing categories, such as Specialty Food, and successful regions, such as the West Pacific, can inform more strategic investment decisions. Plus, producers of Shark Tank could use these insights to curate episodes featuring diverse and high-potential pitches.

Strategically, entrepreneurs should tailor their pitches to highlight strengths in regional and categorical dynamics while balancing financial asks with equity offers. Investors are encouraged to account for interaction effects, such as the synergy between category and location, to make informed decisions and mitigate biases. Future studies could incorporate additional data, such as audience engagement metrics or social media trends, to enhance predictive accuracy further. Expanding the analysis to later seasons may also uncover evolving trends and provide more comprehensive insights. By leveraging these findings, entrepreneurs can refine their pitch strategies, investors can make data-driven decisions, and the competitive funding environment on Shark Tank can become more optimized for success.

Appendix



Appendix Figure 1: PCA Biplot Depicting Relationships Between Metrics and Clusters



Appendix Figure 2: Dominant Principal Component Characteristics by Metrics

The PCA analysis suggests four groups, with the first four principal components explaining nearly 90% of the variance:

- **PC1**: Moderate Asked For ($237,477), higher Requested Equity (17.68%), broad regional distribution (4.9 regions), high Frequency (67.9), and high Unique Sharks (7.14).
- **PC2**: Higher Asked For ($366,857), highest Requested Equity (18.20%), narrower regional distribution (3.8 regions), lower Frequency (32.5), and high Unique Sharks (7.0).

- **PC3**: Lower Asked For ($194,036), moderate Requested Equity (16.81%), moderate regional distribution (4.2 regions), moderate Frequency (35.0), and moderate Unique Sharks (6.73).
- **PC4**: Lowest Asked For ($189,500), lowest Requested Equity (13.44%), narrowest regional distribution (2.8 regions), lowest Frequency (20.0), and high Unique Sharks (7.0).

Appendix Table A: Optimal hyperparameters setting for considered models

Model	Hyperparameter	Considered values
Logistic regression	C (regularization strength)	0.08
	Penalty	L2
	Solver	Glmnet
Classification Tree	cp (complexity parameter)	0.03
Random Forest	mtry	2
	splitrule	Gini
	min.node.size	5
	solver	ranger
XGBoost	nrounds	100
	max_depth	5
	eta (learning rate)	0.01
	gamma	1
	colsample_bytree	0.8
	min_child_weight	10
	subsample	0.7