NYU

Shark Tank US Data Analysis and Prediction

Subhrajit Dey (sd5963@nyu.edu)

Raunak Choudhary (rc5553@nyu.edu)

Saniya Gapchup (syg2021@nyu.edu)

CS-GY 6923 Machine Learning

Contents

01

NYU

Introduction

‘Did you know that only **0.05%** of startups receive venture capital funding? Yet, these few companies drive a significant portion of innovation and economic growth in our economy …’

- The startup ecosystem is critical for innovation, contributing to job creation and technological progress.
- Early-stage funding is a significant challenge for entrepreneurs.
- ***Shark Tank US*** provides a real-world view of investment decision-making by venture capitalists and angel investors.

Research Focus

- Analyzing factors influencing funding decisions on *Shark Tank US*.
- Studying the investment patterns of key investors like Mark Cuban and Barbara Corcoran.
- Bridge gaps in qualitative and quantitative funding assessments.

Goals

- Analyze "shark" investment patterns (industries, amounts).
- Compare predictive models (SVC, Logistic Regression, XGBoost, Random Forest).
- Address data challenges like class imbalance and non-numerical fields.

DataSet Overview

Dataset Introduction:

- Dataset from Kaggle includes 1,360+ pitches and 53 features spanning all 16 seasons of Shark Tank US.

Investment Analysis by Sharks

- Mark Cuban leads with 249 deals, followed by Lori Greiner with 217 deals.
- Largest total Investment by Mark Cuban **$62.9M**, followed by Lori Greiner with **$46.5M.**

Top Industries by Shark

- **Food & Beverage:** Barbara Corcoran, Mark Cuban, Guest Sharks
- **Lifestyle/Home:** Lori Greiner, Kevin O'Leary
- **Fashion/Beauty**: Robert Herjavec, Daymond John

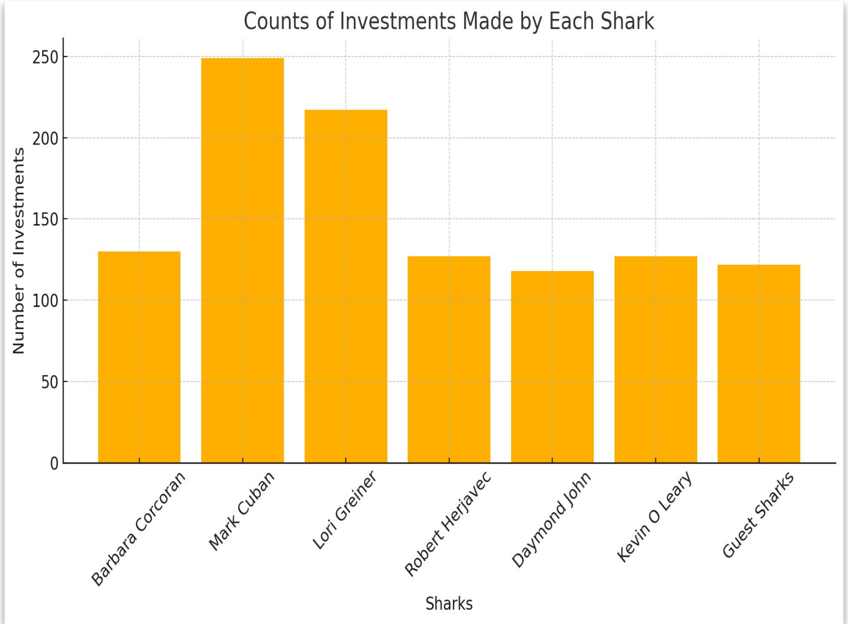
Dataset Fields Overview

- **Key Pitch Details:** Industry, Business Description, Original Ask Amount, Offered Equity etc.
- **Pitcher Specific Information**: Gender, City/State, Multiple Entrepreneurs, Pitcher's age, city etc.
- **Shark specific Data:** Investments by individual sharks (e.g., Barbara Corcoran Investment Amount, Daymond John Equity).

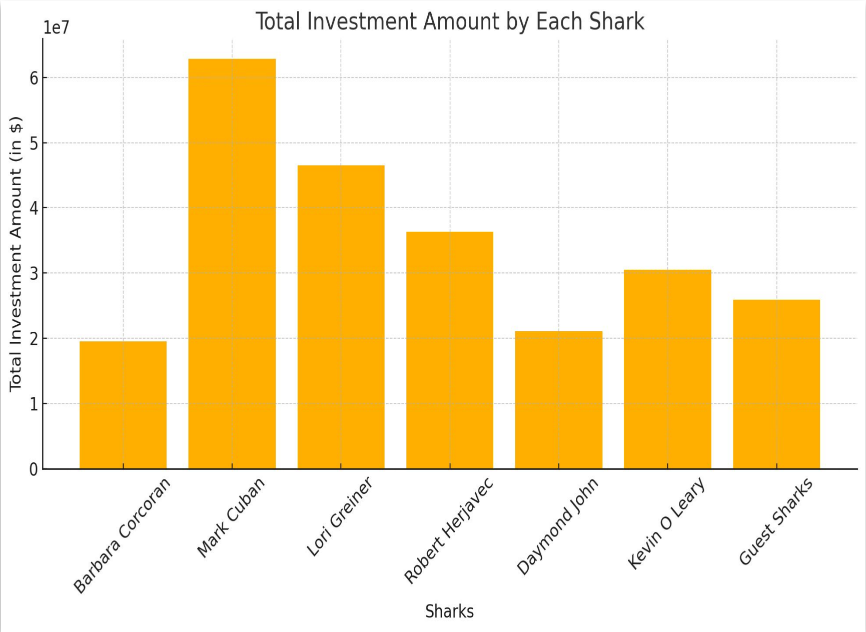
NYU

Investment Charts

- Counts of Investment made by each Shark

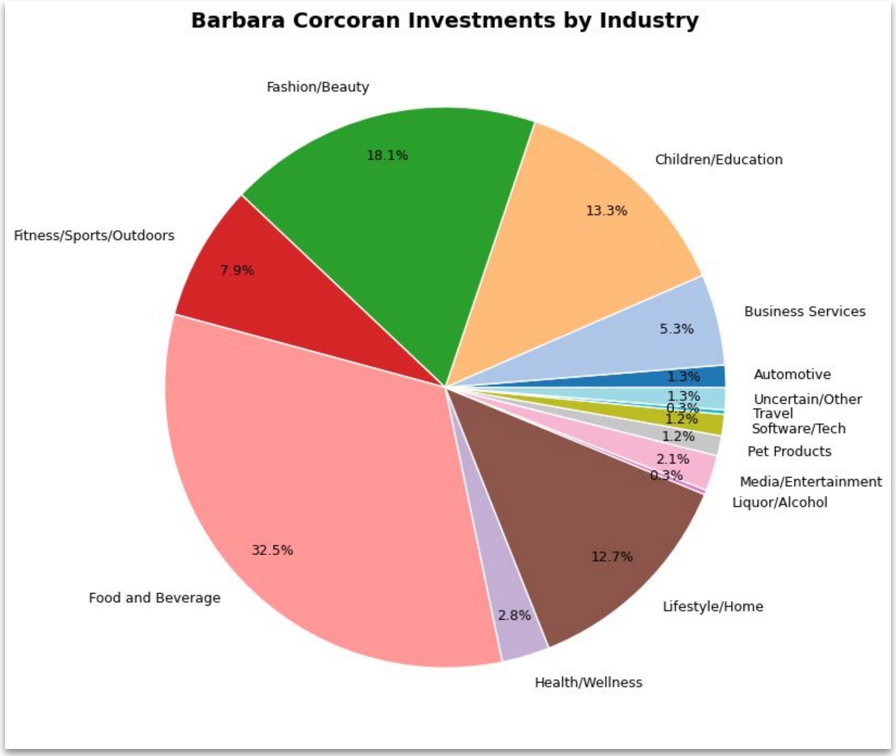


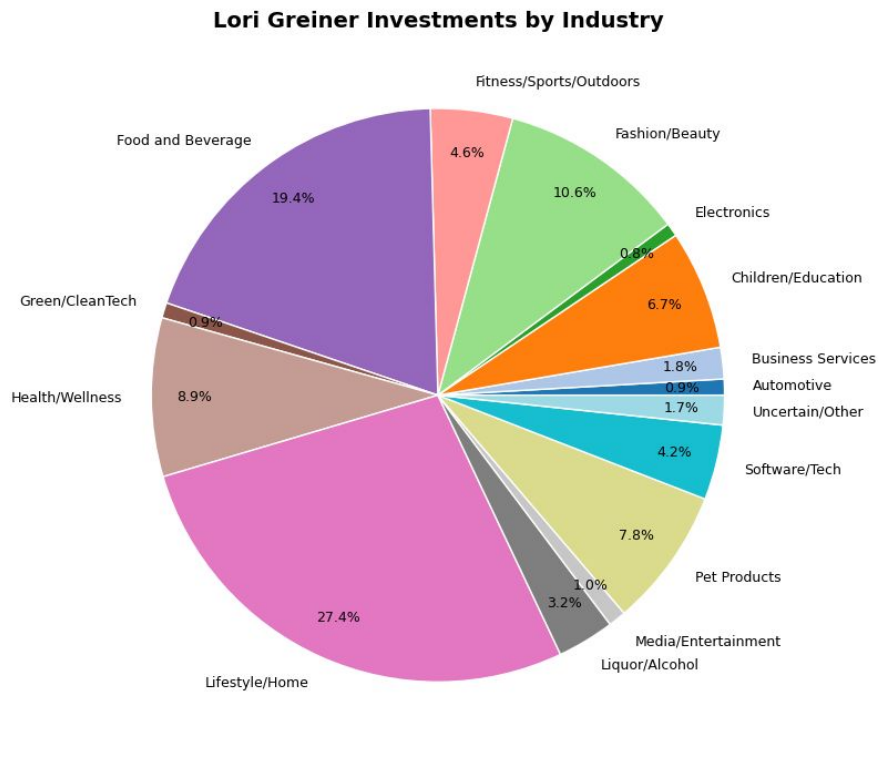
- Total Investment amount by each Shark



NYU

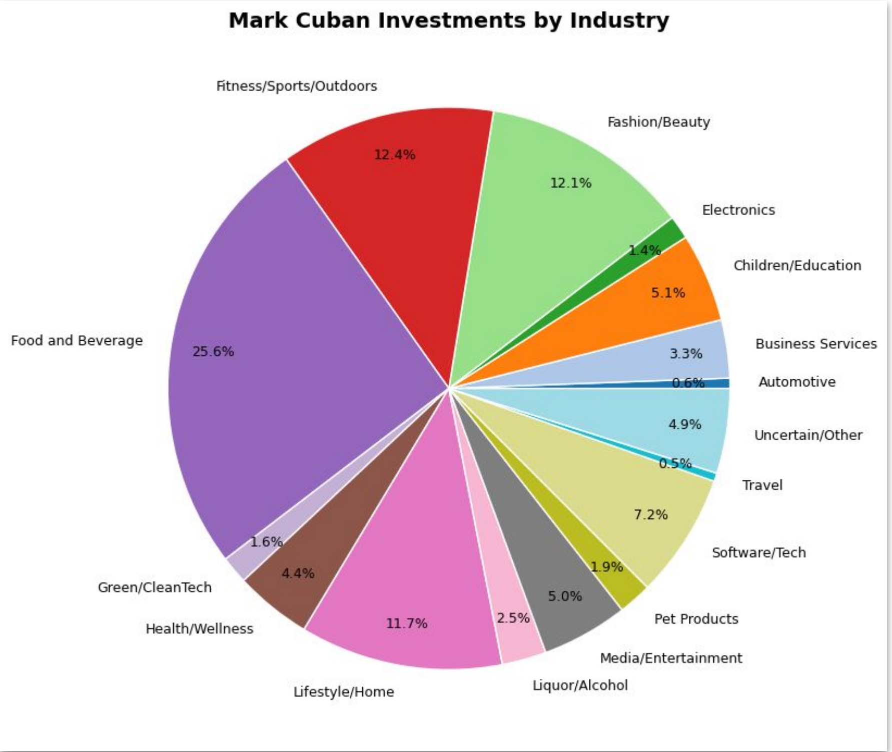
Investment Charts

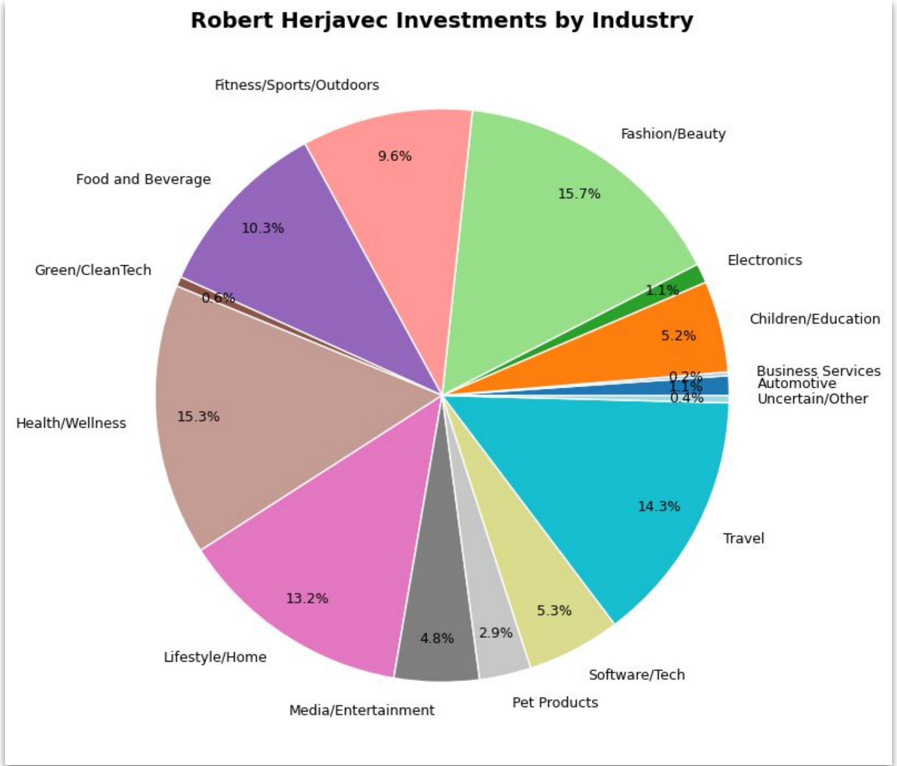




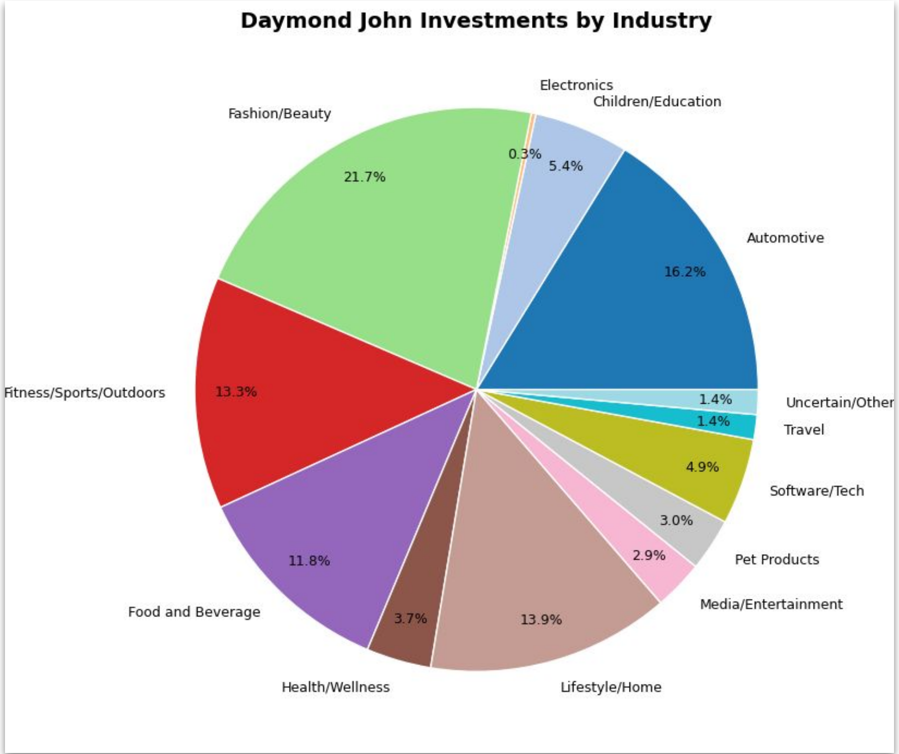
NYU

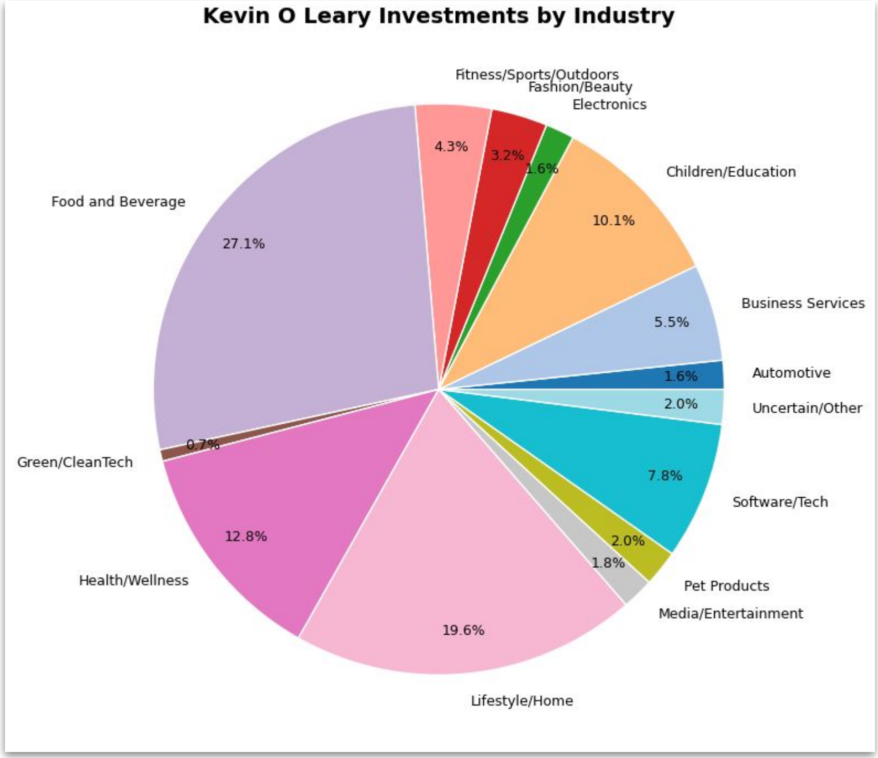
Investment Charts





Investment Charts





NYU

DataSet Preprocessing

Missing Values:

- Numerical fields (Original Ask Amount) imputed with medians.
- Categorical fields (Pitchers Gender) replaced with placeholders.
- Dropped columns with excessive missing data (Pitchers City).

Data Balancing:

- Applied SMOTE to address class imbalance in Investment vs. No Investment.

Feature Engineering:

- Converted text fields (Industry, Business Description) into numerical embeddings via OpenAI Embedding Model.

Embeddings for Textual Data

Why Use Embeddings?

- **Purpose**: Transform textual data into numerical vectors to capture semantic and contextual information.

Applications:

- Identify nuanced relationships in text features like Industry and Business Description.
- Understand patterns influenced by demographic data (e.g., Pitchers Gender).

Benefits of Embeddings

- Enable models to leverage complex, latent relationships in the data.
- Capture subtle patterns that are strong predictors of investment decisions.


3D Word Embeddings Visualization (Top Terms)
Categories
Startup Name
Industry
Business Description
Pitchers Gender

Model Selection and Implementation

1. Logistic Regression (Baseline model)

Definition:

- Statistical model for binary classification.
- Predicts probabilities of binary outcomes using a logistic (sigmoid) function.

Key Concepts:

- Relationship between input features and log-odds.
- Decision threshold (commonly 0.5) for classification.

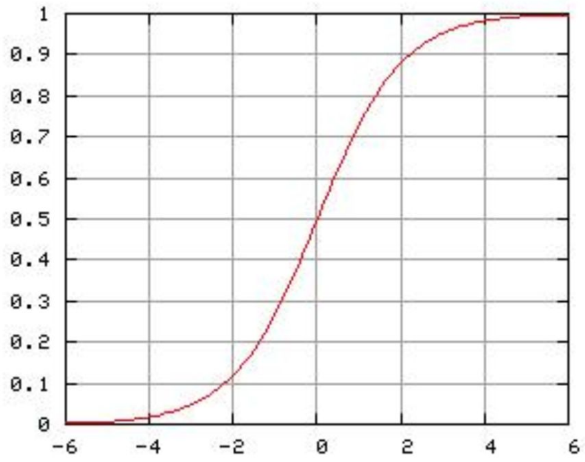


Figure 2: Logistic Regression Curve

Model Selection and Implementation

2. Support Vector Machines (SVM)

- **Definition**:
  - Supervised learning model for classification tasks.
  - Focuses on finding an optimal hyperplane that maximizes the margin between different classes.
- **Key Concepts**:
  - **Support Vectors**: Data points closest to the hyperplane.
  - **Margin**: Distance between the hyperplane and support vectors, maximized for robustness.
- **Handling Non-linear Data**:
  - Uses kernel functions (e.g., linear, polynomial, RBF, sigmoid) to map input into a higher-dimensional space for non-linear decision boundaries.

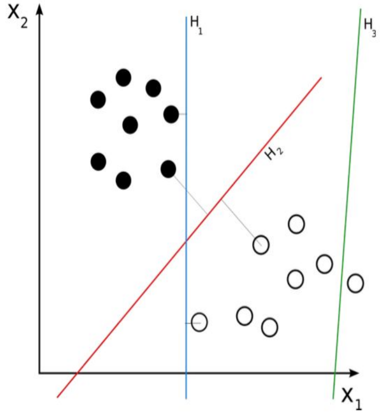


Figure 1: SVM Separating Hyperplanes

Model Selection and Implementation

3. XGBOOST

Definition:

- An advanced machine learning algorithm based on the gradient boosting framework.
- Efficient for both classification and regression tasks, offering scalability and high performance.

Key Concepts:

- **Gradient Boosting**: Iterative process where each tree corrects the errors of the previous trees.
- **Regularization**: L1 and L2 techniques reduce overfitting.
- **Sparse-aware Learning**: Handles missing values effectively.
- **Parallelization**: Fast training via parallelized tree construction.

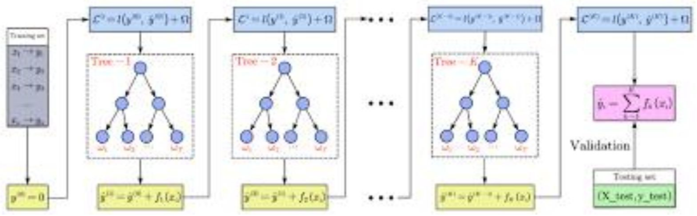


Figure 3: Schematic Diagram of XGBoost

Process Highlights:

- Starts with a base model (e.g., constant value).
- Each tree minimizes residual errors using gradients of the loss function.
- Aggregates tree outputs for final predictions.

Model Selection and Implementation

4. Random Forest

Definition:

- Ensemble learning algorithm combining multiple decision trees for classification and regression.
- Uses majority voting for classification or averaging for regression.

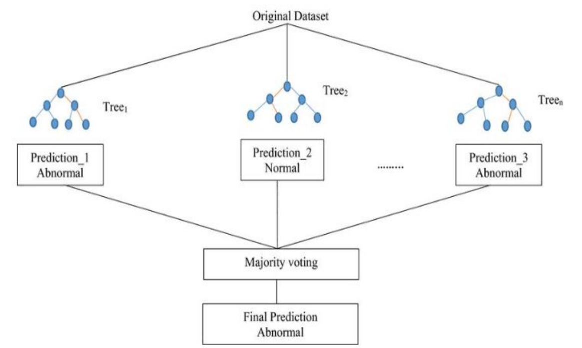


Figure 4: Architecture Diagram of Random Forest

Key Concepts:

- **Randomness**: Trains trees on random subsets of data and features to reduce overfitting.
- **Robustness**: Effective for high-dimensional data and missing values.

Process Highlights:

- Trains each tree independently on bootstrapped subsets of data.
- Aggregates predictions across all trees for final output.

NYU

Hyperparameter Tuning

Steps Implemented:

- **Objective**: Optimize model parameters to improve prediction of shark investments.
- **Approach**: Conducted hyperparameter tuning for Random Forest, Logistic Regression, and XGBoost on validation data using GridSearchCV.
- **SVM:** Performed manual tuning due to its time-intensive nature.

Key Results:

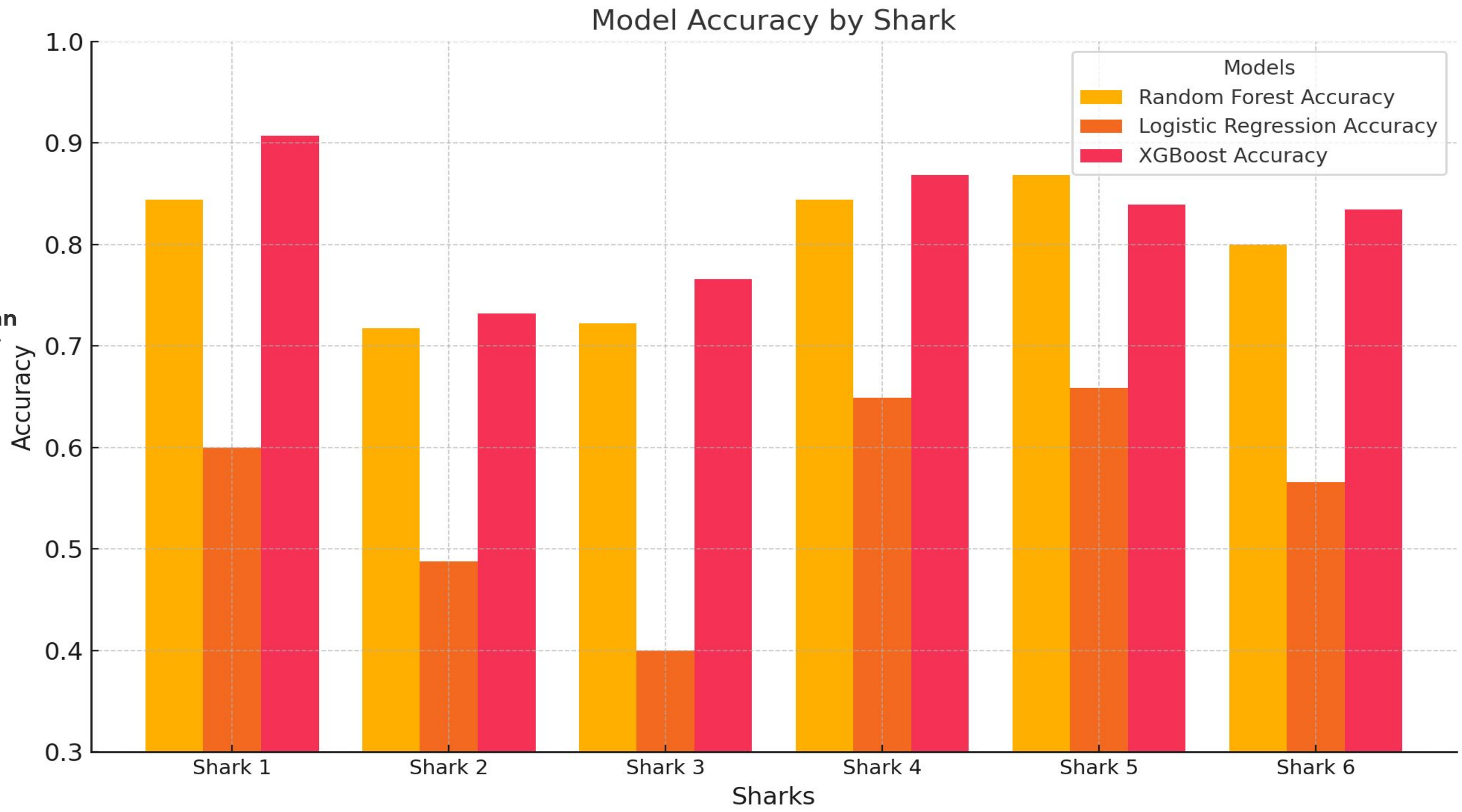
- **XGBoost:** learning_rate=0.1, max_depth=6-9, n_estimators=70-100.
- **Random Forest**: max_depth=None, n_estimators=50-100.
- **Logistic Regression:** Limited performance (<0.66 accuracy) due to dataset complexity**.**
- **SVM:** ‘rbf’ kernel

GridSearchCV:

- **Purpose:** Optimized model parameters to maximize validation accuracy using 5-fold cross-validation.
- **Key Parameters tuned**: **(Random Forest)**: n_estimators, max_depth, **(Logistic Regression)**: C, penalty and **(XGBoost):** learning_rate, max_depth, n_estimators.

Hyperparameter Tuning

Index 1: Barbara Corcoran
Index 2: Mark Cuban
Index 3: Lori Greiner
Index 4: Robert Herjavec
Index 5: Daymond John
Index 6: Kevin O'Leary



NYU

Model training

Classification:

- **Objective:** Train and evaluate models to classify sharks' investment decisions using multi-output binary classification (shark-wise predictions).
- **Models Used:** Random Forest, Logistic Regression, SVC, XGBoost.

Regression:

- **Objective:** Predict exact investment amounts for each shark.
- **Models Used:** RandomForestRegressor and XGBRegressor**.**
- **Key metrics:** RMSE, MSE, MAE and R2 score.

Results: Classification

Random Forest:
- Strengths: Consistently achieved the highest accuracy across sharks.
- Shark 1: 79.02% (Best model), but low recall for "Investment" class (33%).
- Shark 5: Best recall (54%) with an F1-score of 25%, indicating improved minority-class handling.

Logistic Regression:
- Strengths: Served as a baseline but struggled with complex relationships in the data
- Shark 1: 62.44% test accuracy, limited precision and recall for investments.
- Shark 6: Lowest test accuracy at 57.56%, highlighting challenges with imbalanced data.

SVM:
- Performance: Moderate results with the RBF kernel.
- Shark 1: Test accuracy of 66.34%, showing slight improvement over Logistic Regression.
- Shark 5: Achieved 68.78% test accuracy, but computational demands limited scalability.

XGBoost:
- Strengths: Competed closely with Random Forest in accuracy and robustness.
- Shark 1: 75.12% test accuracy, reflecting strong generalization.
- Shark 6: Test accuracy of 74.15%, with balanced performance across classes.

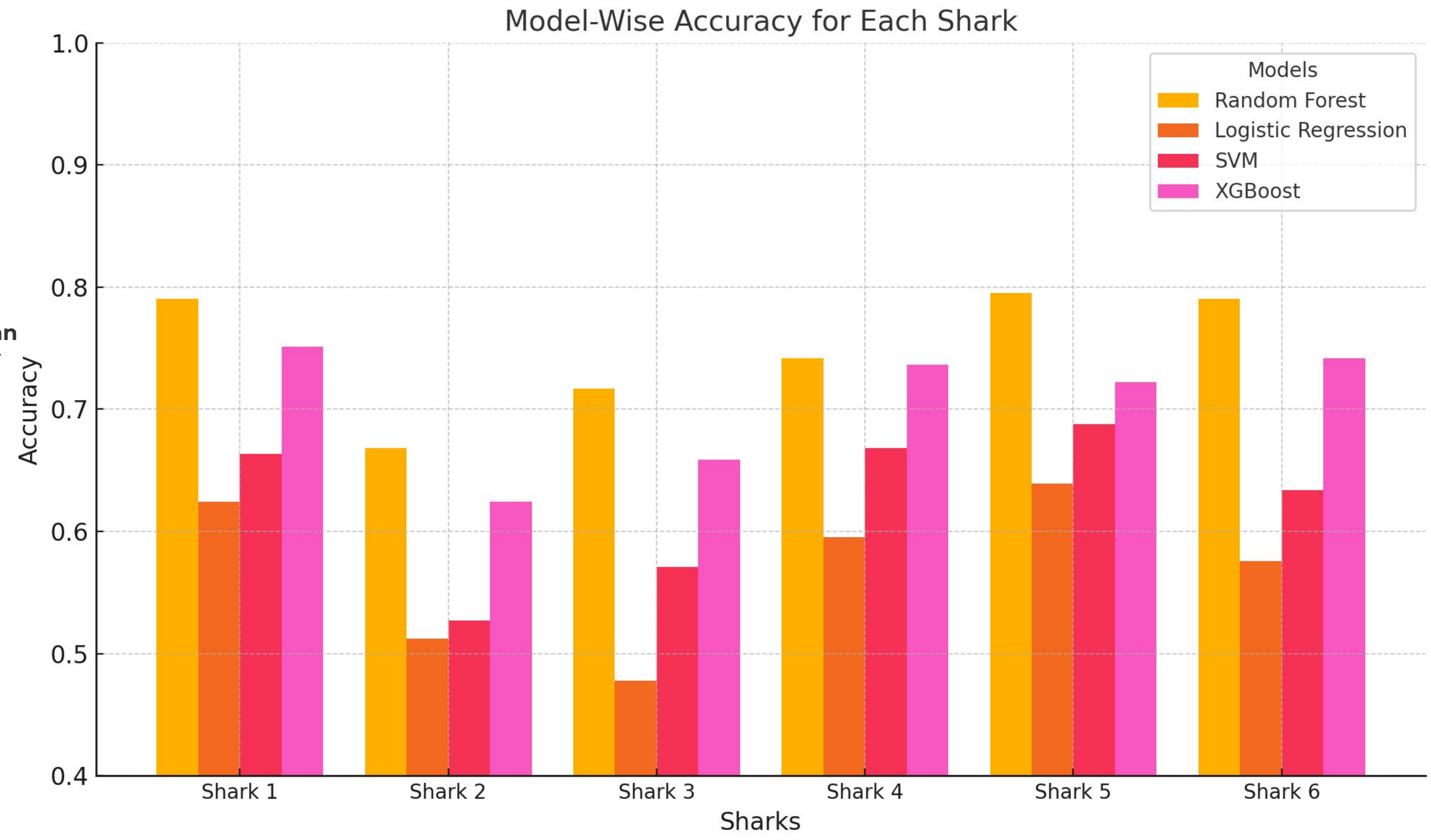
Results: Classification

Shark	Random Forest		Logistic Regression (baseline)		SVM		XGBoost	
	Val	**Test**	**Val**	**Test**	**Val**	**Test**	**Val**	**Test**
Shark 1	**80.49%**	**79.02%**	61.95%	62.44%	66.83%	66.34%	74.63%	75.12%
Shark 2	**67.32%**	**66.83%**	49.76%	51.22%	54.15%	52.68%	60.49%	62.44%
Shark 3	**69.27%**	**71.71%**	40.98%	47.80%	52.68%	57.07%	72.68%	65.85%
Shark 4	**80.98%**	**74.15%**	64.88%	59.51%	63.90%	66.83%	79.51%	73.66%
Shark 5	**82.44%**	**79.51%**	63.41%	63.90%	63.41%	68.78%	74.63%	72.20%
Shark 6	**75.12%**	**79.02%**	56.59%	57.56%	63.90%	63.41%	71.71%	74.15%

NYU

Results: Classification

Index 1: Barbara Corcoran
Index 2: Mark Cuban
Index 3: Lori Greiner
Index 4: Robert Herjavec
Index 5: Daymond John
Index 6: Kevin O’Leary



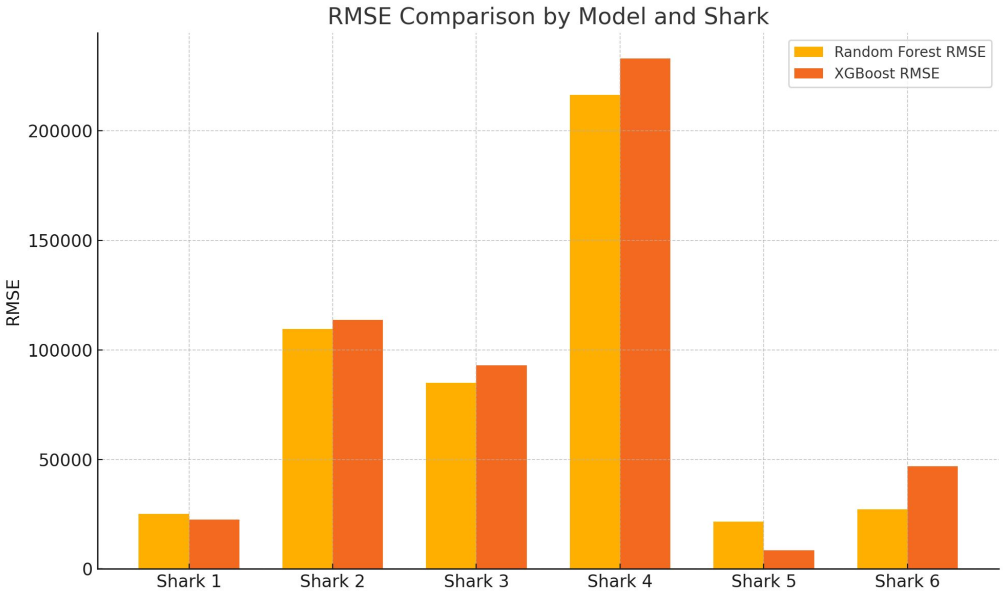
NYU

Results: Regression

Key insights:

- **XGBoost:** Best model for Shark 1 (RMSE: 22570.87, R^2: 0.8353) and Shark 5 (RMSE: 8524.34, R^2: 0.9578).
- **Random Forest:** Best for Shark 2 (RMSE: 109626.95, R^2: 0.5898), Shark 3 (RMSE: 85036.95, R^2: 0.6453), and Shark 6 (RMSE: 27258.33, R^2: 0.8910).
- **Challenges:** Shark 4 showed poor results with Random Forest (RMSE: 216530.50, R^2: -5.0121), likely due to outliers and high variability.

Index 1: Barbara Corcoran
Index 2: Mark Cuban
Index 3: Lori Greiner
Index 4: Robert Herjavec
Index 5: Daymond John
Index 6: Kevin O'Leary



Results: Regression

Shark	Model	Test RMSE
Shark 1	Random Forest	24,994.13
	XGBoost	**22,570.87**
Shark 2	**Random Forest**	**109,626.95**
	XGBoost	113,666.52
Shark 3	**Random Forest**	**85,036.95**
	XGBoost	93,039.88
Shark 4	**Random Forest**	**216,530.50**
	XGBoost	233,194.26
Shark 5	Random Forest	21,579.29
	XGBoost	**8,524.34**
Shark 6	**Random Forest**	**27,258.33**
	XGBoost	46,840.22

NYU

Challenges and Overcoming them

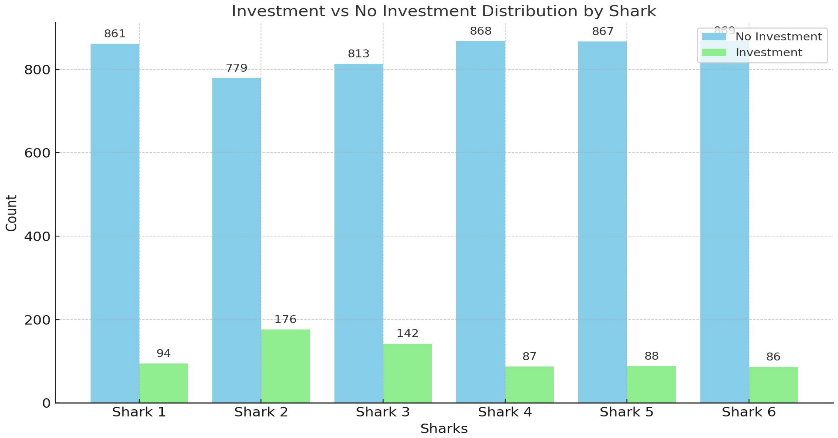
Non-Numerical Data:

- **Problem**: Features like Business Description and Startup Name posed challenges in direct analysis.
- **Solution**: Used OpenAI embeddings to convert text into numerical representations, followed by PCA for dimensionality reduction.

High Dataset Imbalance:

- **Problem**: Imbalance in investment ("minority class") vs. non-investment data affected model performance.
- **Solution**: Applied SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic data points and balance classes.

Index 1: Barbara Corcoran
Index 2: Mark Cuban
Index 3: Lori Greiner
Index 4: Robert Herjavec
Index 5: Daymond John
Index 6: Kevin O'Leary



Real World Applications

Investment Insights:

- Analyzing shark-specific investment patterns can help aspiring entrepreneurs tailor their pitches to align with individual investor preferences.
- Predictive models provide valuable insights into the likelihood of securing investments, empowering startups to refine their strategies.

Interactive Online Game:

- Create a Shark Tank Simulation Game, allowing users to pitch their ideas to virtual AI sharks modeled after real investors.
- Competitive Element: Friends can compete to secure the highest investment, fostering creativity and entrepreneurial thinking.
- Educational Value: Gamifies the startup funding process, helping users understand the dynamics of pitching and investor decision-making.

Thank you