CONTENTS

Business Problem

Securing funding is a pivotal challenge for emerging companies in the growing startup ecosystem. This project addresses the critical need for a predictive model to accurately assess the likelihood of a startup receiving funding, facilitating a more efficient and informed decision-making process for investors and entrepreneurs.

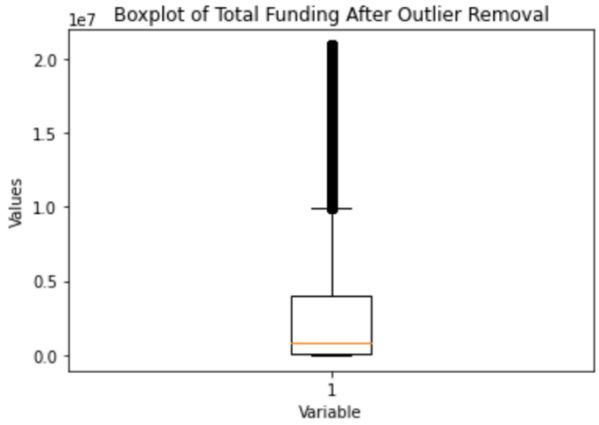
Background/History

The startup funding world has changed a lot, with more money available but also more competition. Although there's plenty of funding, not all startups find it easy to get the money they need due to the complex investment process. This highlights the importance of tools that can predict if a startup will get funding. Such tools can improve how startups devise their strategies and how investors choose where to allocate their resources.

Data Explanation

The foundation of this analysis is the "Startup Investments Crunchbase" dataset, a comprehensive compilation of data from approximately 54,000 startups. The dataset includes information on funding rounds, investors, industry classifications, and other pertinent details.

- **Data Preparation:** The initial phase involved cleaning and preprocessing the dataset to ensure data quality and consistency. This included correcting inconsistencies, adjusting

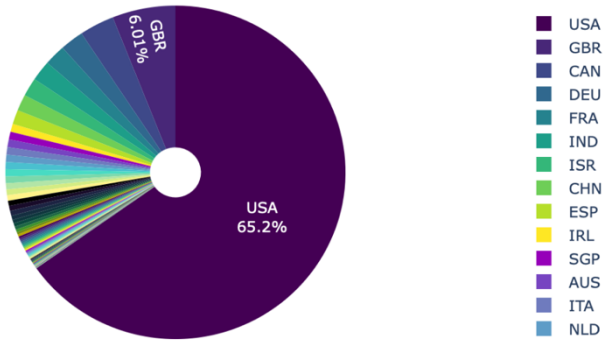


Explanation:

Even thought not completely removed, the before and after box plot shows a significant removal of outliers using IQR.

data types, eliminating duplicates, and handling null values. Outliers in total funding were identified and managed using the Interquartile Range (IQR) method, supplemented by visual analysis to maintain data integrity.

- **Data Dictionary Highlights:**
  - Funding Rounds: The number of times a startup has successfully secured funding.
  - Venture Funding: The amount of venture capital a startup has received.
  - Startup Market: The startup's industry segment operates within, condensed into 43 distinct categories.
  - Age at First Funding: The time elapsed from a startup's inception to its first round of funding, with an average of 40.57 months.

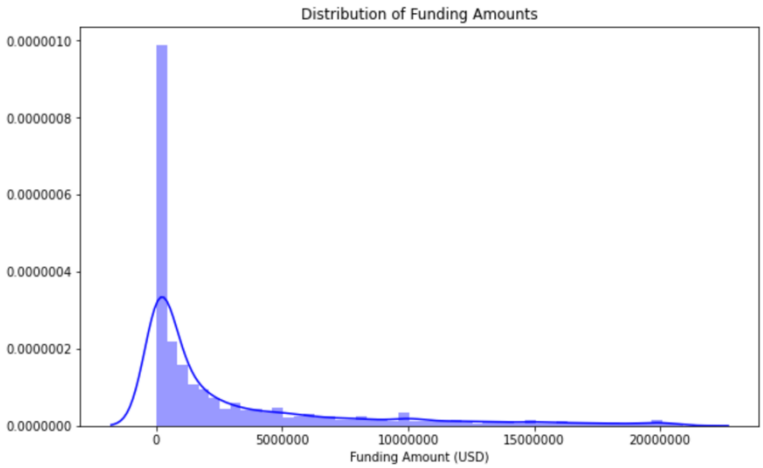


Methods

Logistic Regression and Decision Tree models were utilized due to their proven effectiveness and interpretability in classification tasks. Performance evaluation encompassed accuracy, precision, recall, and F1 score, employing cross-validation to ensure the models' generalizability.

Analysis

- **Exploratory Data Analysis (EDA):** Key findings from the EDA phase included an average duration of 40.57 months to first funding, with a prevalent funding range of 1-10 million dollars.



The analysis also highlighted the dominance of sectors like software, biotechnology, and e-commerce and a geographical concentration of startups in the United States, followed by the United Kingdom and Canada.

- **Feature Engineering:** This phase was crucial for transforming and creating features that enhance the model's predictive capability. Consolidating 709 market values into 43 distinct industry groups significantly optimized the dataset for analysis.

- **Feature Selection:** The selection process was guided by insights from EDA and correlation analysis, identifying critical predictors of funding status such as funding rounds, venture funding, and startup market.

Conclusion

Model	Accuracy	Precision	F1 Score	Confusion Matrix
Logistic Regression	0.8358	0.8473	0.9091	[[84, 846], [93, 4695]]
Decision Tree Classifier	0.7899	0.8717	0.9091	[[311, 619], [582, 4206]]

The comparative analysis revealed that Logistic Regression outperformed the Decision Tree model, achieving an accuracy of 83.58% and a precision of 84.73%. Both models registered an F1 score of 90.91%, underscoring their effectiveness in predicting funding status. However, Logistic Regression demonstrated a higher efficiency with fewer misclassifications for non-funded startups, making it the preferred model for predicting startup funding likelihood.

Assumptions

The analysis assumes that the dataset accurately reflects the broader startup ecosystem and that the features selected directly influence the likelihood of a startup receiving funding.

Limitations

The study's limitations are tied to the dataset's scope and the inherent complexity of the startup funding process. The predictive power of the models is contingent on the dataset's representation of the real-world scenario, which may not capture all variables influencing funding decisions.

Challenges

Addressing missing data, managing outliers, and ensuring model accuracy without overfitting represented significant challenges. Moreover, selecting predictive features required a deep dive into the dataset, balancing statistical significance with practical relevance.

Future Uses/Additional Applications

Future research could explore integrating alternative machine learning models, such as Support Vector Machines (SVM) and neural networks, to refine predictions. Additionally, incorporating features related to the startup team, product offerings, and marketing strategies could provide a more granular analysis of funding likelihood.

Recommendations

Based on the findings, we recommend implementing the Logistic Regression model for stakeholders in the startup ecosystem seeking to predict funding outcomes. This model balances accuracy and interpretability, which is essential for practical application.

Implementation Plan

We propose continuously validating the Logistic Regression model against new and diverse datasets to operationalize the findings. This involves iterative training to adapt to market changes, integrating the model into investment analysis tools, and developing a user-friendly interface for startups and investors to evaluate funding prospects.

Ethical Assessment

Ethical considerations include.

- ensuring the privacy and security of startup data,
- mitigating biases in model training and predictions, and
- transparently communicating the limitations and uncertainties of predictive outcomes.

Creating a space where predictive analytics boosts fairness, diversity, and innovation in startups is crucial. Depending on whether a startup has funding, a decision tree classifier might be preferable. If reducing false positives is the priority, logistic regression could be a better fit. The choice of model should be tailored to each specific situation, considering all important aspects.

GitHub

https://github.com/rozank/dsc680_applied_datascience/tree/main/Project-1

References

Andy_M. (2020, February 17). *Startup investments (crunchbase)*. Kaggle. https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase?resource=download

What industries are included in Crunchbase?. What Industries are included in Crunchbase? (n.d.). https://support.crunchbase.com/hc/en-us/articles/360043146954-What-Industries-are-included-in-Crunchbase-

Questions

Here is the list of 10 possible questions that audience might ask.

1. What makes the "Startup Investments Crunchbase" dataset suitable for this analysis?
2. Why were Logistic Regression and Decision Tree models chosen for this study?
3. How did you handle missing data and outliers in your dataset?
4. What are the critical predictors of a startup's funding status that you determined from this project?
5. Why did Logistic Regression outperform the Decision Tree model?
6. Can you explain the significance of the models' F1 scores?
7. What are the limitations of your study?
8. How did you ensure the models' predictive power and avoid overfitting?
9. What ethical considerations were considered in your analysis?
10. Why is model interpretability important in your study, and how does it affect the choice between Logistic Regression and Decision Trees?