**Analysis of Startup Funding and its Success**

**Summary:**

Startups struggle to get funding onto their table, which is the main reason for their success or failure. It is interesting to know the failure story of a startup, similar to the success story. Understanding the investment rounds, capital burning, and the type of investors the startups get on board. The business can be impacted by the infrastructure and the location where it will operate, so it is essential to choose the company's location while choosing the firm.

Such crucial factors are essential for a business to understand what is lacking and where to improve. We aim to achieve these critical insights from the available dataset. Finally, for a given location and category, we need to predict if the market is ready to accept that business and make the startup successful.

We have extracted the data related to the startup companies and their expansion over time, funding rounds, and acquisitions. We are curious to know the percentage of acquisitions happening for each startup and of which how many are turning into unicorns and gaining success. The datasets contain three tables: investments, companies and acquisitions. It comprises over 60k companies established from the late 1970s to mid-2010s, which can be described as shown below:

● Acquisitions - This table details the companies and their acquired companies.

● Additions - This table details the types of rounds and the number of companies in each round for a given year.

● Investments - This table details the companies and their investors and the number of investments raised for different rounds.

● Companies - This table has information about startup companies.

**Project Goals:**

We want to perform Exploratory Data Analysis on the data to drive insights, which are mentioned below:

1. Distribution of funding raised over a time frame.

2. Companies in focus: Companies that have raised the most funding.

3. Establishing a relationship between the industries and the feasibility of a business

based on the location.

4. Investment firms in focus: Investment companies that can redirect funds into the

startup economy.

5. Amount raised through different funding rounds like venture capital, seed

investment etc.

6. Significant acquisitions made by a company created by an acquirer.

7. How long does a startup take to be acquired?

We have worked on data tidying by removing null values, bringing the data to a standardised format and imputing data wherever possible for the model to have enough training data. Transformed and standardised data when needed for processing. Based on the data provided, we have used prediction models to predict if a startup will likely succeed.

**Methods:**

As a part of this project, we have used various data tidying, transformation, visualisation and modelling techniques.

A separate section is written for the methods involved with technical descriptions.

**Results:**

We have modelled the data using 4 four different models, but the CAT Boost algorithm with oversampling technique has proven to be the most effective, with an accuracy of 76.4 per cent and an AUC score of 75.8 per cent.

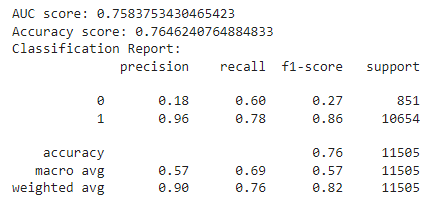
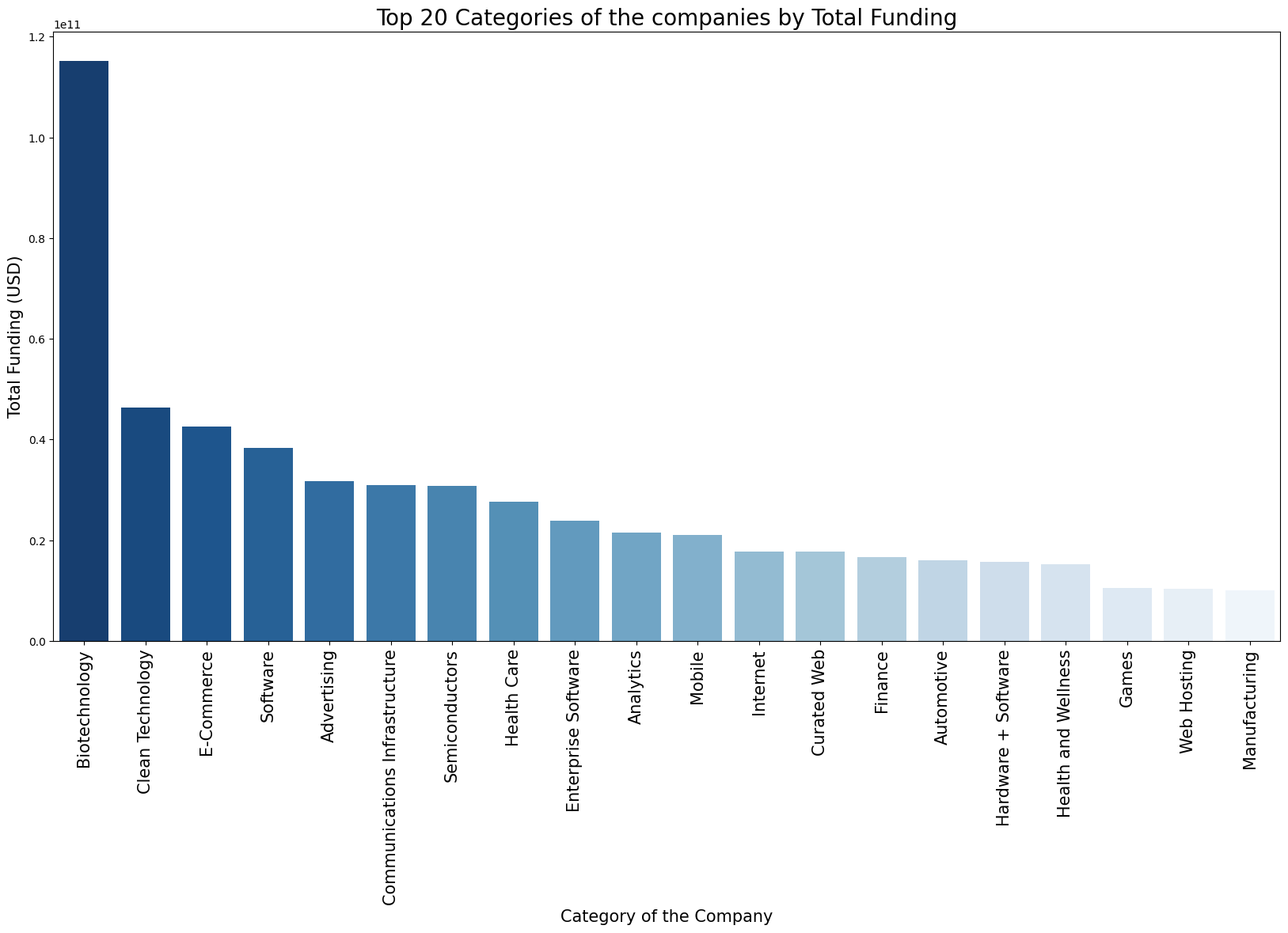
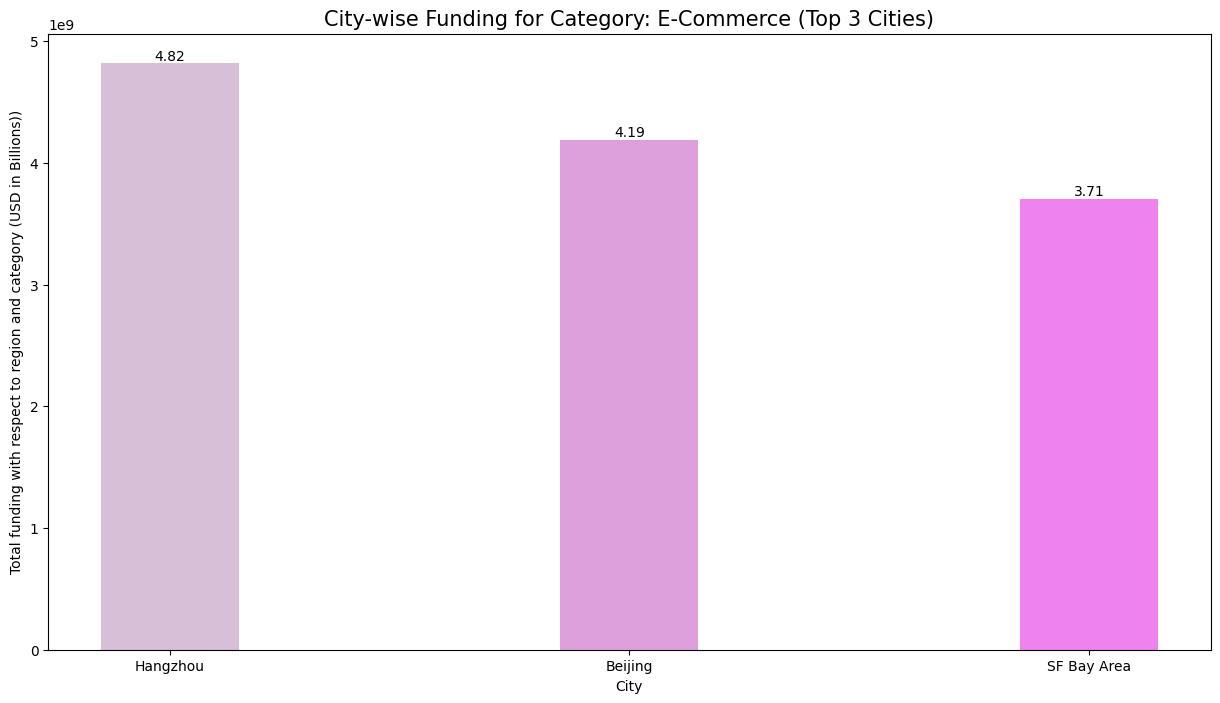
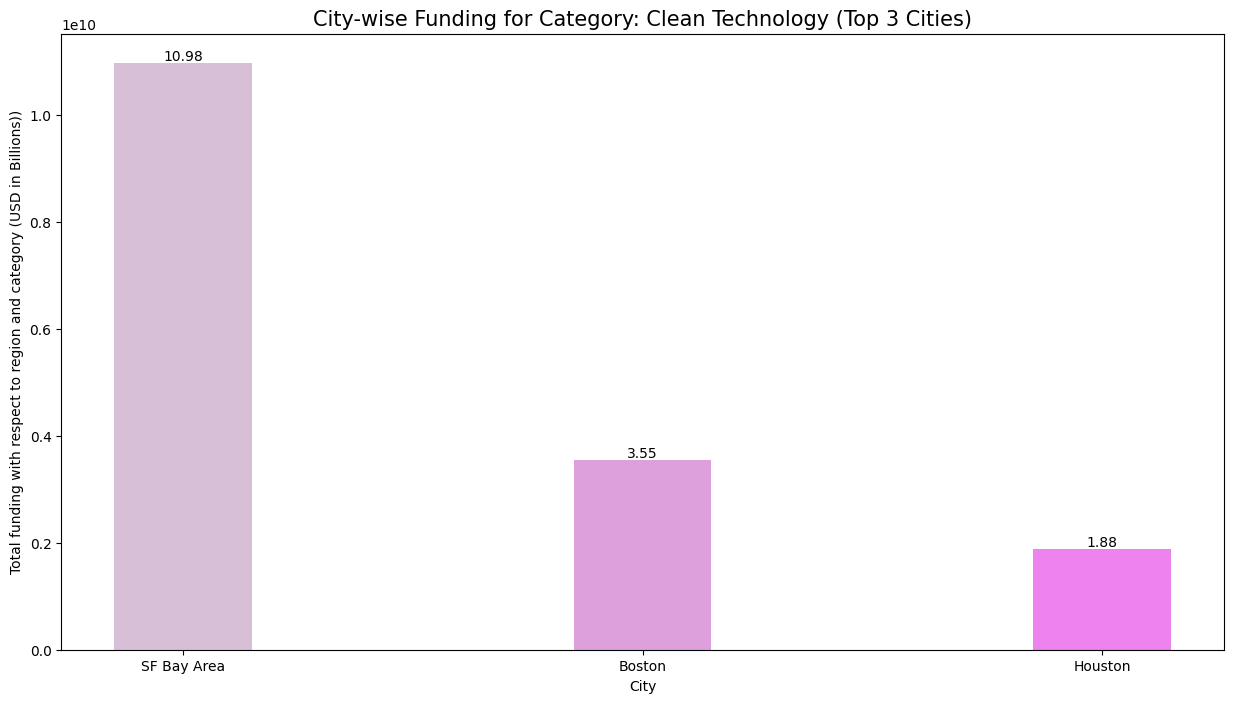
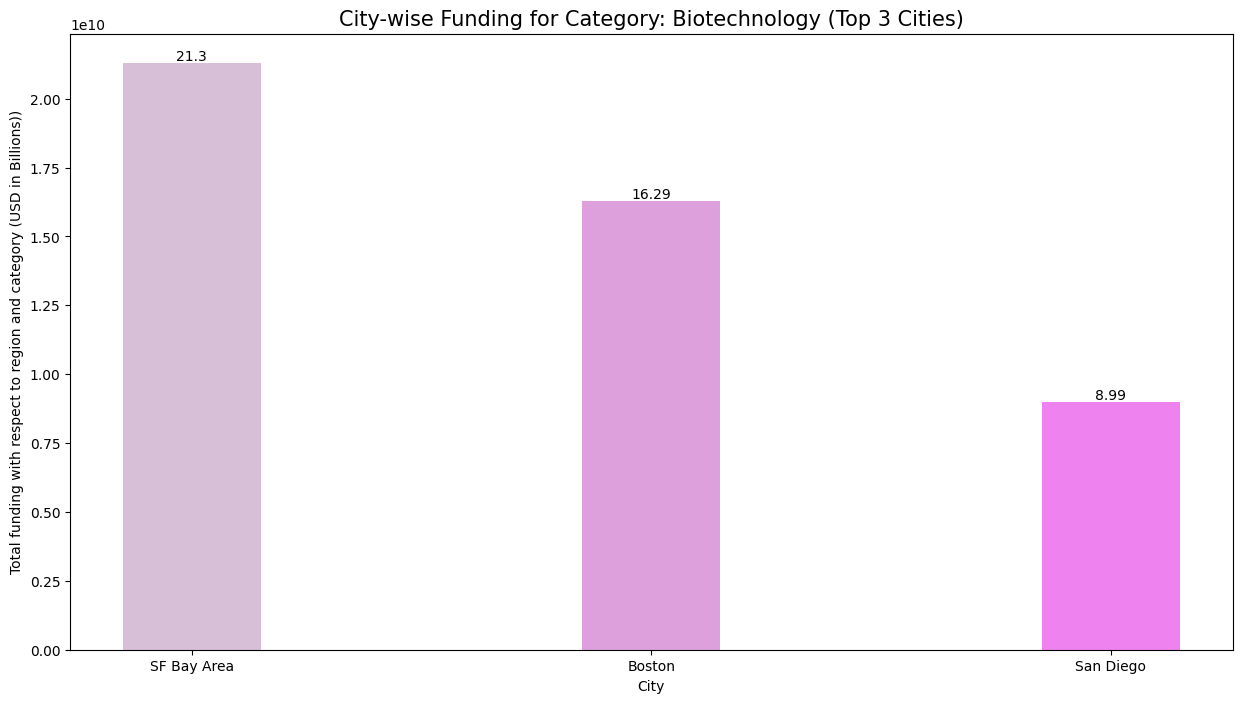


Fig. Model parameters for CAT Boost Algorithm with oversampling technique

We have drawn a few inferences from the Exploratory Data Analysis and included some of them here.

The graph below depicts the Top 20 categories of companies that have received the most funding, out of which Biotechnology, Clean Technology and E-commerce have received the most funding. We will also look at the top regions for each category mentioned above, from which we can see that the cities below are hubs for companies belonging to specific categories.

****

****

**Discussion:**

Our project will help new startup entrants decide if their startup will succeed based on various inferences drawn from the data and model prediction; location and ease of funding will be the most important factors to be considered, and our project will help ease off this process.

This project can be improved by continuously integrating newly available data of startups and feeding this entire model into a website, making it available for the general public to predict the success of their startups.

**Statement of Contributions:**

This project is a collaborative effort amongst the four of us, and each of us contributed equally to the timely completion of this project. However, we are listing out the major contribution of each individual towards a specific category.

Nikhileshwar Reddy Bommareddy: Data Cleaning, Model building and EDA

Vinesh Kumar Gande: Data Cleaning, Model Building and EDA

Lahari Boni: Data Cleaning, EDA and Reporting

Sai Akhil Rayapudi: Data Cleaning, Reporting and Model Building

**References:**

Data Source

*GitHub - notpeter/crunchbase-data: 2015 CrunchBase Data Export as CSV.*

*GitHub.* [*https://github.com/notpeter/crunchbasedata*](https://github.com/notpeter/crunchbasedata)

FIPS Data

*FIPS data extracted from -* [*http://efele.net/maps/fips-10/data*](http://efele.net/maps/fips-10/data)

**Appendix:**

Model building code:

CAT Boost- Oversampling

*# Split dataset into training and testing sets*

*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)*

*# Apply oversampling to the training setros = RandomOverSampler(random\_state=42) X\_train\_resampled, y\_train\_resampled = ros.fit\_resample(X\_train, y\_train)*

*# Train catboost model*

*cat\_model = CatBoostClassifier(iterations=1000, learning\_rate=0.1, random\_seed=42) cat\_model.fit(X\_train\_resampled, y\_train\_resampled, verbose=False)*

*# Evaluate the model using accuracy, AUC, and classification*

*reporty\_prob = cat\_model.predict\_proba(X\_test)[:, 1]*

*auc\_score = roc\_auc\_score(y\_test, y\_prob)*

*print('AUC score:', auc\_score)*

*# Calculate accuracy*

*scorey\_pred = cat\_model.predict(X\_test)*

*acc\_score = accuracy\_score(y\_test, y\_pred)*

*print('Accuracy score:', acc\_score) print("Classification Report:") print(classification\_report(y\_test, y\_pred))*

Link to the entire code book:

<https://github.com/LahariBoni/Startup-Success-Analysis>