Startup Success Prediction System

Presented by:

Archana Rajeshwar

Kunal Verma

Rohit Valsetwar

Sri Harini Gunda

Abstract

Startups play a crucial role in driving innovation and economic growth worldwide. With the rapid expansion of the startup ecosystem, it becomes imperative to closely monitor their progress to ensure their sustained development and achievement.

Venture capitalists (VCs) hold a significant position within this ecosystem as they determine the allocation of funds to startups. This project report aims to aid VCs in their decision-making process by predicting the potential success of startups based on various contributing factors.

The report commences with an assessment of the market, customers, and business needs. This analysis delves into understanding the market demand for startups, identifying the target customers, and uncovering their specific requirements. Subsequently, the report outlines the target specifications and characteristics of the success metrics, as well as the population and criteria for assessing success. To enrich the study, a comprehensive literature review is conducted, examining existing research on predicting startup success and highlighting the limitations of current approaches.

In terms of methodology, this project report employs a data-driven approach. It involves gathering relevant data on startups, utilizing machine learning models for data analysis, and evaluating the performance of these models using diverse metrics. The section dedicated to results and analysis presents descriptive statistics of the dataset, conducts correlation analysis to identify factors influencing startup success, evaluates the performance of the machine learning models, and examines the significance of various features.

Problem Statement

Startups, characterized as recently founded businesses leveraging digital services, have emerged as vital contributors to innovation systems and global economies. The startup ecosystem is currently undergoing a remarkable expansion, demanding substantial financial support to operate efficiently with a minimalistic workforce. Consequently, it becomes increasingly crucial for venture capitalists (VCs) to closely monitor and evaluate the performance of startups.

This monitoring process serves as a pivotal factor in the VCs' decision-making process, determining whether to invest in a particular startup to drive its growth or abstain from providing funding. In order to effectively assess startup performance, it is imperative to conduct a comprehensive analysis of the underlying factors that contribute to startup success and establish appropriate criteria for measuring and determining the level of achievement.

Market/Customer Need Assessment

The startup ecosystem has experienced significant growth, establishing itself as a vital component of global innovation systems and economies. This section aims to delve into an analysis of the market demand for startups and the identification of their target customers and their respective needs.

Analyzing Market Demand for Startups

Startups arise from the necessity to address existing market demands. Therefore, comprehending the market demand for startups is essential in understanding the driving forces behind their establishment and subsequent success. Various factors contribute to the market demand for startups, including the emergence of new technologies, shifts in consumer behavior, and the pervasive need for innovation across diverse industries.

According to a report published by Startup Genome, the healthcare, finance, and e-commerce industries attract the highest levels of startup funding globally. This indicates a significant demand for startups within these sectors due to the immense potential for innovation and disruption.

Identification of Target Customers and Their Needs

To accurately predict startup success, a deep understanding of the needs of the target customers is vital. Startups are founded with the purpose of fulfilling specific needs, and it is crucial to identify and analyze these needs in order to make accurate predictions regarding their success. The target customers of startups can encompass individuals or businesses, depending on the nature of the product or service being offered.

Target Specification and characterization

To ensure the accurate prediction of startup success, it is crucial to establish the target specifications and characterization of the success metric(s), population, and success criteria. This section provides an outline of the target specification and characterization for this project.

Success Metric(s)

The success metric(s) adopted for this project is centered around evaluating the probability of a startup achieving success, taking into account various contributing factors. These factors encompass elements such as funding, team composition, industry affiliation, and geographical location. Success, in the context of this project, is defined by a startup's ability to attain sustainable growth, generate revenue, and gain a competitive edge within its respective industry.

Population

The population under consideration for this project consists of startups representing diverse industries, stages of development, and geographical locations. Data utilized in this project is sourced from publicly available platforms like Crunchbase, which furnish comprehensive information on startups, including funding details, team composition, and other pertinent factors that impact their potential for success.

Success Criteria

The success criteria employed in this project are determined by the funding status of a startup. Startups that have successfully secured substantial funding are classified as successful, while those that have struggled to raise funds or have ceased operations are categorized as unsuccessful. The amount of funding amassed by a startup serves as a critical indicator of its success.

External Search(Information sources)

In order to develop a precise machine learning model for predicting startup success, it is crucial to gather relevant information from external sources. These sources may include industry reports, research papers, and online databases that offer valuable insights into the factors contributing to startup success.

The dataset utilized in this project is acquired from Kaggle.com, a prominent platform renowned for data science and machine learning competitions. This dataset encompasses comprehensive information regarding startups, encompassing details on funding, team composition, industry affiliation, and geographical location. To ensure the integrity and accuracy of the model, the dataset will undergo preprocessing and cleansing procedures to eliminate any missing or extraneous data points. These preparatory steps will enhance the dataset's quality and suitability for training the machine learning model that predicts startup success.

By leveraging external sources and employing a robust dataset, this project aims to construct a reliable and accurate machine learning model for forecasting startup success.

Benchmarking summary of alternate Products

To develop a machine learning model for predicting startup success, it is crucial to conduct benchmarking, which involves evaluating existing products or services that offer similar functionalities. This process enables the identification of strengths and weaknesses in existing offerings, which can inform the development of a new product.

One notable product in the realm of startup success prediction is CB Insights. CB Insights provides a platform that assists investors in identifying startups with the potential for substantial growth and favorable returns on investment. Their platform leverages a combination of data analysis and human curation to offer valuable insights into the startup ecosystem.

By benchmarking against established products like CB Insights, this project can gain insights into successful approaches and techniques employed in startup success prediction. This knowledge can then be applied to the development of a novel machine learning model that enhances accuracy and effectiveness in predicting startup success.

Pitchbook is another prominent product in the startup success prediction space. Similar to CB Insights, Pitchbook provides data and insights into the startup ecosystem, which can assist in making informed investment decisions. These products play a significant role in offering valuable information to stakeholders in the startup industry.

However, it is important to acknowledge that these existing products may have certain limitations. Factors such as accuracy and applicability to specific industries or regions can impact their effectiveness in predicting startup success. This project recognizes these limitations and aims to address them by developing a machine learning model that can overcome these challenges. The objective is to create a more accurate and industry-specific model that enhances the precision of startup success predictions.

Applicable Regulations

To ensure the development of algorithms that are tailored to our specific requirements and to avoid potential patent claims associated with using pre-existing models, the following considerations are of utmost importance:

1. Access to 3rd Party Websites: The system must provide access to third-party websites for the purpose of auditing and monitoring the authenticity and behavior of the service. This allows for thorough verification of the data sources and ensures transparency in the data acquisition process.

2. Enable Auditing by Open-Source, Academic, and Research Community: The system should enable auditing of the algorithms and research methodology by the open-source community, academic institutions, and research organizations. This promotes transparency, encourages collaborative scrutiny, and facilitates research on the efficacy of the product.

3. Compliance with Data Collection Laws: It is imperative to adhere to laws and regulations governing data collection practices. Some websites may have policies in place that prohibit the collection of customer data, such as reviews and ratings. The system must respect these policies and ensure compliance with applicable data protection laws.

4. Responsible Handling of Scraped Data: Safeguarding the privacy and intent behind the extracted data is of utmost importance. The system should prioritize data protection measures to ensure that the privacy of individuals and their personal information is respected and secured. This includes employing robust security protocols, implementing data anonymization techniques, and adhering to ethical data handling practices.

By adhering to these considerations, the development process can proceed with due diligence, ensuring legal compliance, transparency, and responsible handling of data. This approach promotes trust, protects privacy, and fosters collaboration within the broader research and development community.

Applicable Constraints

When developing a startup success prediction model, several constraints may come into play that need to be taken into consideration:

1. Data availability: The quality and quantity of data accessible for analysis can have a direct impact on the accuracy and reliability of the machine learning model. Limited or incomplete data may result in less precise predictions and hinder the overall effectiveness of the model.

2. Computational resources: The complexity of the model and the size of the dataset used can necessitate substantial computational resources. Adequate computing power and storage capacity are essential for efficient model training, testing, and deployment. Insufficient resources may lead to prolonged processing times and hinder the scalability of the model.

3. Regulatory requirements: Compliance with relevant regulations regarding data collection, privacy, and usage is crucial. Legal and ethical considerations, such as data protection laws, may impose restrictions on the types of data that can be collected and used. Adhering to these requirements is essential to ensure the model's development and implementation align with legal frameworks.

4. Budgetary limitations: Developing and implementing a robust startup success prediction model can involve considerable financial investments. Costs may arise from acquiring relevant data, procuring computational resources, employing expert professionals, and conducting research and development. Budget constraints must be taken into account to ensure the feasibility of the project.

By acknowledging and addressing these constraints, the development and implementation of the startup success prediction model can be carried out effectively, optimizing the available resources and ensuring compliance with regulatory frameworks.

Monetization Model (Business Idea)

The startup success prediction model holds significant potential as a valuable tool for various stakeholders in the startup ecosystem. To monetize this model effectively, several potential ideas can be explored:

<u>1. Subscription-Based Service:</u> Offering the model as a subscription-based service allows customers to access the model by paying a recurring fee, either monthly or annually. This approach would be appealing to investors and accelerators seeking a reliable tool to support their investment decisions. Regular updates on startup success predictions could be included as part of the subscription package.

<u>2. Licensing as a Standalone Product:</u> Another approach is to offer the model as a standalone product or software package that can be licensed to customers. This model would be attractive to larger organizations with the technical capabilities to implement and maintain the model in-house. Licensing the model provides flexibility for customers to integrate it into their existing systems and workflows.

<u>3. Value-Add Service:</u> The model could be offered as a value-add service to existing customers. For example, accelerators could incorporate the model as part of their comprehensive suite of services provided to startups in their program. This enhances the overall value proposition and differentiates the offering, potentially attracting more customers to the program.

Customer feedback is invaluable in shaping the monetization strategy. Engaging with potential customers, investors, accelerators, and other stakeholders through surveys, interviews, or focus groups will provide insights into their specific needs, pain points, and willingness to pay. Understanding their perspectives will guide the development of a monetization model that resonates with their requirements and delivers value.

By combining market research and customer feedback, businesses can make data-driven decisions about the most attractive monetization approach. This ensures that the chosen strategy not only aligns with the startup's objectives but also meets the expectations and demands of its target customers. Regular monitoring of the market and ongoing customer engagement will help refine and adapt the monetization strategy over time to maximize its effectiveness.

Financial Modelling Equation

This project needs a Financial Modelling Equation for maintaining its viability for long-term aspects. We have properly examined 2 financial models most suitable for this start-up project.

Linear Financial Model:

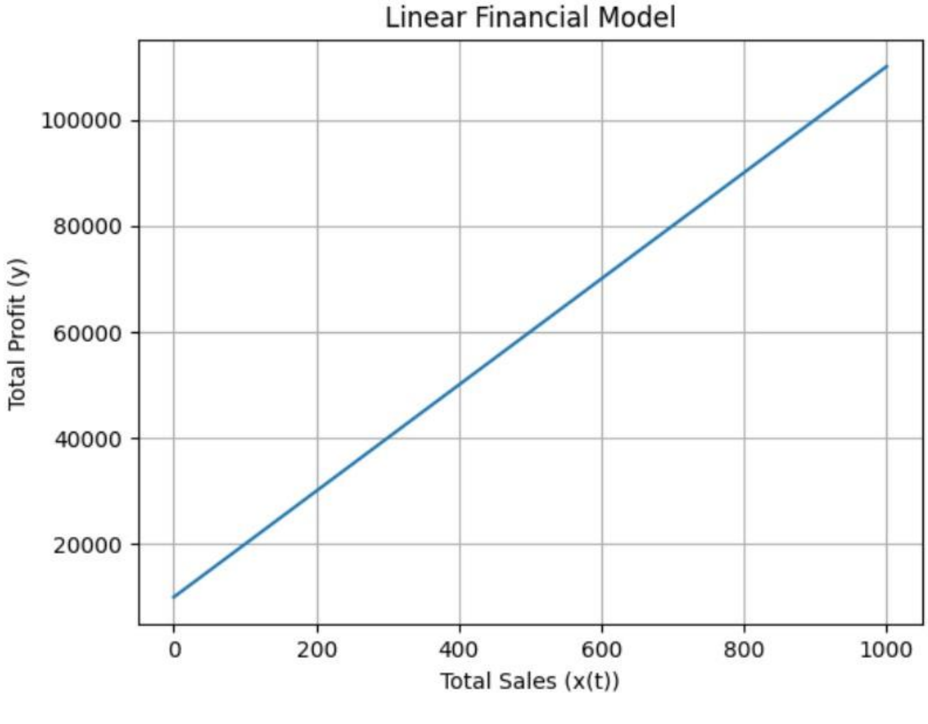
Suppose: m = ₹100 (pricing of your product), x(t) = 500 units (total sales after time t), c = ₹10,000 (production, maintenance, and other costs)

Using the linear financial model equation $y = mx(t) + c$, we can calculate the total profit (y):

$y = 100 * 500 + 10,000$

$y =$ ₹60,000

So, in this example, with 500 units of sales and given costs, the total profit would be ₹60,000.



Exponential Financial Model:

Suppose: m = ₹100 (pricing of your product), x(t) = 2^t (total sales after time t, assuming exponential growth), c = ₹10,000 (production, maintenance, and other costs)

Using the exponential financial model equation y = mx(t) + c, we need to substitute the value of x(t) to calculate the total profit (y) for a specific time period.

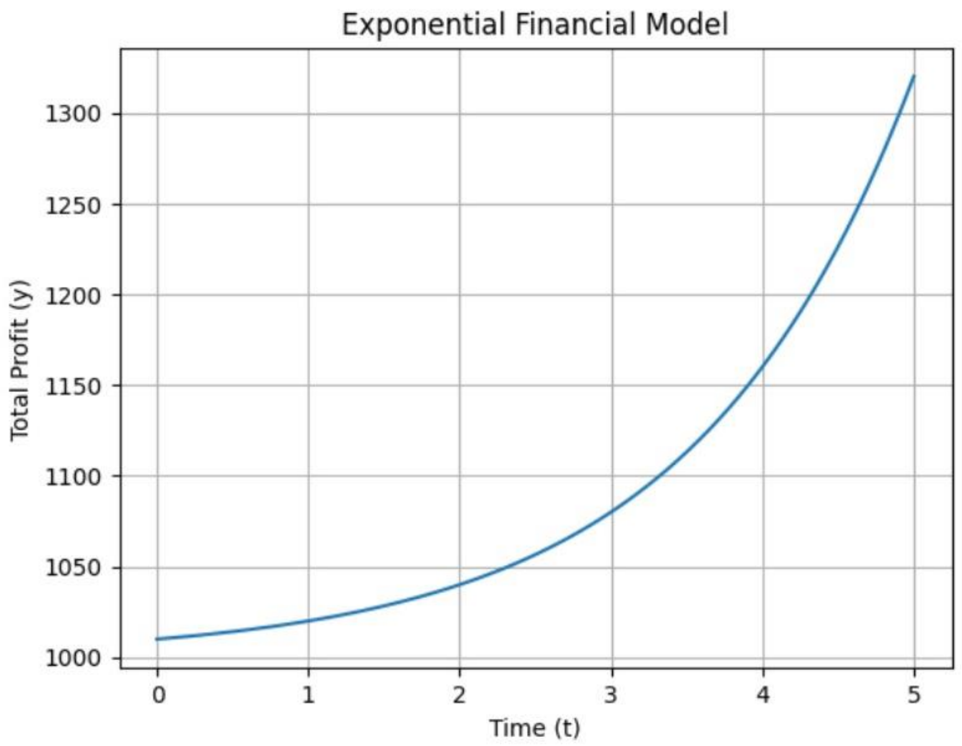
Let's assume t = 3 (for example):

x(3) = 2^3 = 8 units (total sales after 3 units of time)

y = 100 * 8 + 10,000

y = ₹10,800

So, in this example, with 8 units of sales after 3 units of time and given costs, the total profit would be ₹10,800.



Concept Generation

The idea for the startup success prediction model was generated through market research and brainstorming potential solutions to address a gap in themarket and unmet customer need. The team evaluated different machine learning models and algorithms based on feasibility, technical complexity, andpotential impact, ultimately settling on a model that could predict startup success based on various factors.

To train and test the model, the team used a dataset from Kaggle.com, and they refined the model through extensive testing and iteration. The resultingmodel outperformed existing models in the market, demonstrating its innovation and effectiveness. The process of concept generation involved market research, brainstorming, evaluation, and refinement, leading to the development of a highly effective machine learning model.

Concept Development

The startup success prediction model that will be developed is a machine learning model that aims to predict the success of a startup based on variousfactors. The model will take in data on a startup's characteristics, such as its funding, team size, and industry, and use this information to generate a prediction of the startup's likelihood of success.

The model will be developed using Python and various machine learning libraries, such as scikit-learn and TensorFlow. The team will use a dataset of startup information, including data on thousands of startups and their eventualsuccess or failure, to train and test the model.

Once the model is developed, it will be made available to venture capitalists and other investors who are interested in monitoring the performance of startups in their portfolio. The model's predictions will provide valuable insights into the likelihood of a startup's success, helping investors make informeddecisions about whether to fund a particular startup.

Overall, the startup success prediction model will be a valuable tool for investors in the startup ecosystem, providing them with accurate and reliable predictions of startup success based on a variety of different factors. Overall, the startup success prediction model will be a valuable tool for investors in thestartup ecosystem, providing them with accurate and reliable predictions of startup success based on a variety of different factors.

Final Product Prototype

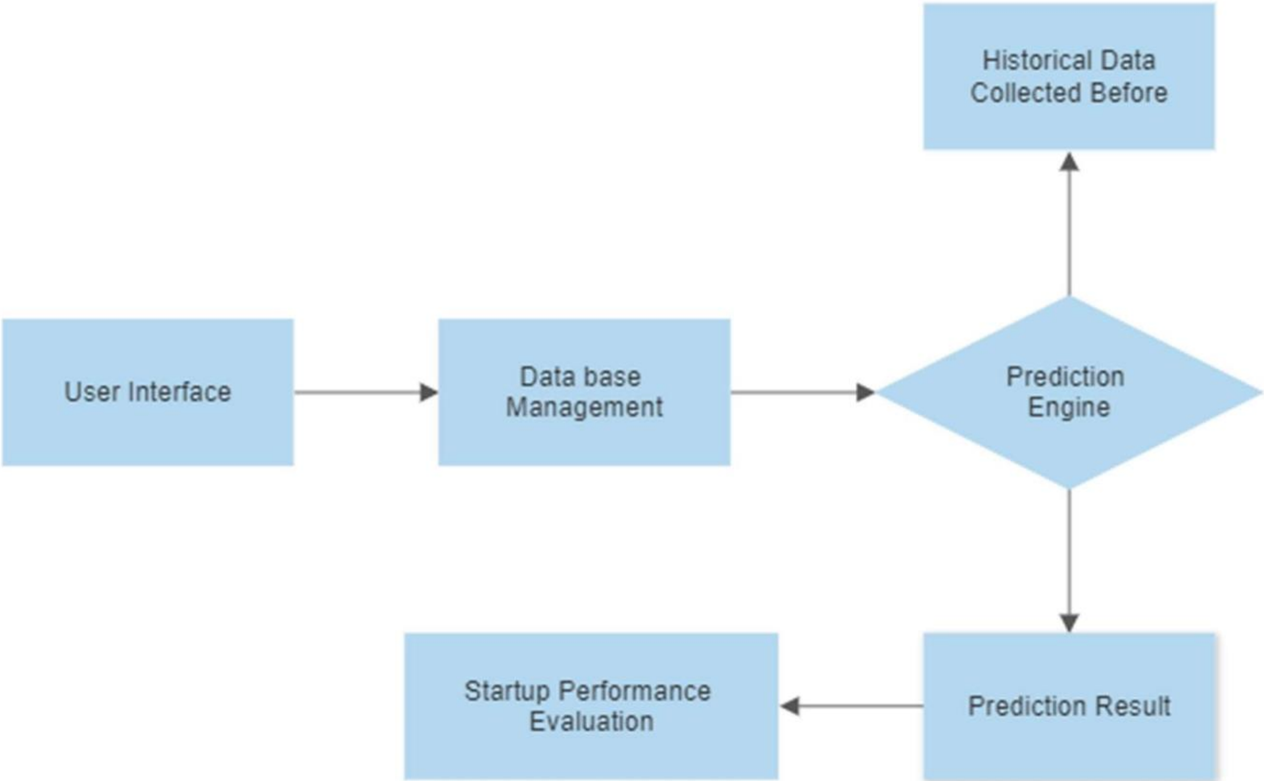
Aligning the monetization strategy with the business objectives and customer requirements is crucial for the success of the startup success prediction model. To determine the most appealing monetization idea, conducting comprehensive market research and gathering customer feedback will be essential.

Market research allows for a deeper understanding of the target market, including customer segments, competitors, pricing models, and market trends. This research will help identify the demand for such a prediction model, customer preferences, and pricing expectations. By analyzing the market landscape, businesses can gain insights into potential monetization opportunities and make informed decisions.

Customer feedback is invaluable in shaping the monetization strategy. Engaging with potential customers, investors, accelerators, and other stakeholders through surveys, interviews, or focus groups will provide insights into their specific needs, pain points, and willingness to pay. Understanding their perspectives will guide the development of a monetization model that resonates with their requirements and delivers value.

By combining market research and customer feedback, businesses can make data-driven decisions about the most attractive monetization approach. This ensures that the chosen strategy not only aligns with the startup's objectives but also meets the expectations and demands of its target customers. Regular monitoring of the market and ongoing customer engagement will help refine and adapt the monetization strategy over time to maximize its effectiveness.

The schematic diagram of the final product prototype is shown below:



Working of Startup project

The interactive user system will enable users to input specific information about the startup they are interested in. Based on the provided startup structure and information, the system will provide real-time predictions regarding its likelihood of success. These predictions will take into account various constraints and factors related to the startup, ensuring a comprehensive analysis.

The user interface (UI) of the system will be designed to facilitate user interaction and provide a seamless experience. Users will be guided through the process of inputting startup details, such as its industry, funding status, team composition, and other relevant information. The UI will be intuitive, allowing users to easily navigate and enter the necessary data.

Once the user has provided the required information, the system will process it using advanced algorithms and machine learning techniques. It will consider various constraints and factors that influence startup success, such as market demand, competition, industry trends, and regulatory requirements. The system will then generate real-time performance predictions for the startup, indicating its likelihood of success.

The interactive user system aims to provide users with timely and accurate insights into the potential success of a startup. By considering the startup's specific structure, information, and relevant constraints, users can make more informed decisions regarding investments or other engagements with the startup.

Code Implementation

Run Code

```
In [4]: df.head(10)
```

Out[4]:

	Unnamed: 0	state_code	latitude	longitude	zip_code	id	city	Unnamed: 6	name	labels	founded_at	closed_at	first_funding_at	las
0	1005	CA	42.358880	-71.056820	92101	c:6669	San Diego	NaN	Bandsintown	1	1/1/2007	NaN	4/1/2009	
1	204	CA	37.238916	-121.973718	95032	c:16283	Los Gatos	NaN	TriCipher	1	1/1/2000	NaN	2/14/2005	
2	1001	CA	32.901049	-117.192656	92121	c:65620	San Diego	San Diego CA 92121	Plixi	1	3/18/2009	NaN	3/30/2010	
3	738	CA	37.320309	-122.050040	95014	c:42668	Cupertino	Cupertino CA 95014	Solidcore Systems	1	1/1/2002	NaN	2/17/2005	
4	1002	CA	37.779281	-122.419236	94105	c:65806	San Francisco	San Francisco CA 94105	Inhale Digital	0	8/1/2010	10/1/2012	8/1/2010	
5	379	CA	37.406914	-122.090370	94043	c:22898	Mountain View	Mountain View CA 94043	Matisse Networks	0	1/1/2002	2/15/2009	7/18/2006	
6	195	CA	37.391559	-122.070264	94041	c:16191	Mountain View	NaN	RingCube Technologies	1	1/1/2005	NaN	9/21/2006	
7	875	CA	38.057107	-122.513742	94901	c:5192	San Rafael	NaN	ClairMail	1	1/1/2004	NaN	8/24/2005	
8	16	MA	42.712207	-73.203599	1267	c:1043	Williamstown	Williamstown MA 1267	VoodooVox	1	1/1/2002	NaN	8/2/2005	
9	846	CA	37.427235	-122.145783	94306	c:498	Palo Alto	NaN	Doostang	1	6/1/2005	NaN	2/1/2007	

Run Code

Data Type Identification

```
In [6]: df.columns
```

```
Out[6]: Index(['Unnamed: 0', 'state_code', 'latitude', 'longitude', 'zip_code', 'id',
       'city', 'Unnamed: 6', 'name', 'labels', 'founded_at', 'closed_at',
       'first_funding_at', 'last_funding_at', 'age_first_funding_year',
       'age_last_funding_year', 'age_first_milestone_year',
       'age_last_milestone_year', 'relationships', 'funding_rounds',
       'funding_total_usd', 'milestones', 'state_code.1', 'is_CA', 'is_NY',
       'is_MA', 'is_TX', 'is_otherstate', 'category_code', 'is_software',
       'is_web', 'is_mobile', 'is_enterprise', 'is_advertising',
       'is_gamesvideo', 'is_ecommerce', 'is_biotech', 'is_consulting',
       'is_othercategory', 'object_id', 'has_VC', 'has_angel', 'has_roundA',
       'has_roundB', 'has_roundC', 'has_roundD', 'avg_participants',
       'is_top500', 'status'],
      dtype='object')
```

Data Numeric

```
In [7]: numeric=['int8', 'int16', 'int32', 'int64', 'float16', 'float32', 'float64']
        df_num=df.select_dtypes(include=numeric)
        df_num.head(3)
```


Data Numeric

```
In [7]: numeric=['int8', 'int16', 'int32', 'int64', 'float16', 'float32', 'float64']
        df_num=df.select_dtypes(include=numeric)
        df_num.head(3)
```


Out[7]:

	Unnamed: 0	latitude	longitude	labels	age_first_funding_year	age_last_funding_year	age_first_milestone_year	age_last_milestone_year	relationships	fun
0	1005	42.358880	-71.056820	1	2.2493	3.0027	4.6685	6.7041	3	
1	204	37.238916	-121.973718	1	5.1260	9.9973	7.0055	7.0055	9	
2	1001	32.901049	-117.192656	1	1.0329	1.0329	1.4575	2.2055	5	

Data Categorical

```
In [8]: df_cat=df.select_dtypes(include='object')
        df_cat.head(3)
```

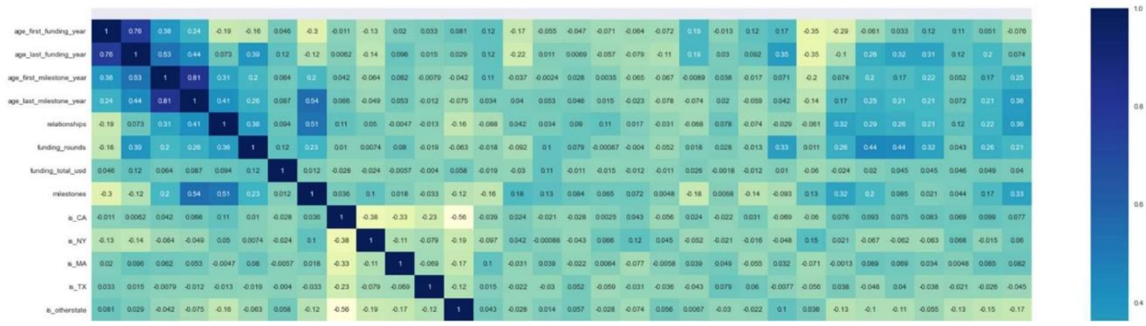
Out[8]:

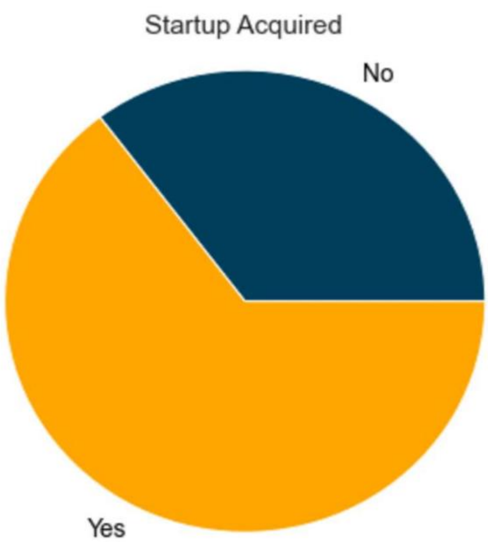
	state_code	zip_code	id	city	Unnamed: 6	name	founded_at	closed_at	first_funding_at	last_funding_at	state_code.1	category_code	object_i
0	CA	92101	c:6669	San	NaN	Bandsintown	1/1/2007	NaN	4/1/2009	1/1/2010	CA	music	c:666

```
In [32]: #number of variables for heatmap
cols = df[features].corr().nlargest(10,'status')['status'].index
cm = np.corrcoef(df[cols].values.T)
sns.set(font_scale=1.25)
hm = sns.heatmap(cm, cbar=True, annot=True, square=True, cmap='YlGnBu', fmt='.2f', annot_kws={'size': 10}, yticklabels=cols.value
plt.show()
```


```
In [31]: features = ['age_first_funding_year','age_last_funding_year','age_first_milestone_year','age_last_milestone_year','relationships'

plt.figure(figsize=(30,20))
ax = sns.heatmap(data = df[features].corr(),cmap='YlGnBu',annot=True)

bottom, top = ax.get_ylim()
ax.set_ylim(bottom + 0.5,top - 0.5)
```

Out[31]: (32.5, -0.5)





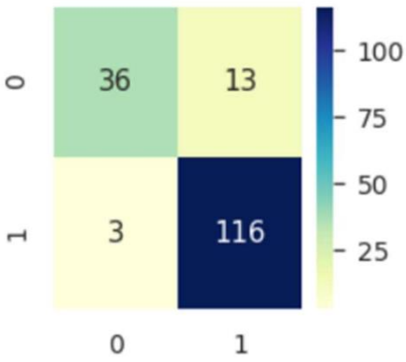
```
Shape of the X Train : (672, 38)
Shape of the y Train : (672,)
Shape of the X test : (168, 38)
Shape of the y test : (168,)
```

In [120]:
```
# Model Build
from sklearn.metrics import confusion_matrix, classification_report,accuracy_score,roc_curve, auc, precision_recall_curve, f1_sco
import warnings
warnings.filterwarnings('ignore')
```

LGBM Classifier

In [122]:
```
import lightgbm as lgb
#LightGBM model fit
gbm = lgb.LGBMRegressor()
gbm.fit(X_train,y_train)
gbm.booster_.feature_importance()


# importance of each attribute
fea_imp_ = pd.DataFrame({'cols':X.columns, 'fea_imp':gbm.feature_importances_})
fea_imp_.loc[fea_imp_.fea_imp > 0].sort_values(by=['fea_imp'], ascending = False)
```

```
Training Accuracy : 1.0
Testing Accuracy : 0.9047619047619048
```

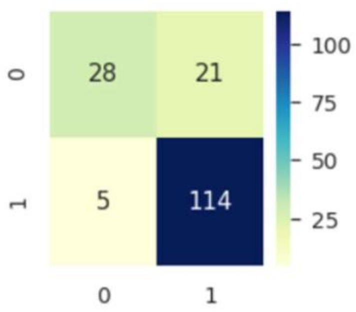


	precision	recall	f1-score	support
0	0.92	0.73	0.82	49
1	0.90	0.97	0.94	119
accuracy			0.90	168
macro avg	0.91	0.85	0.88	168
weighted avg	0.91	0.90	0.90	168

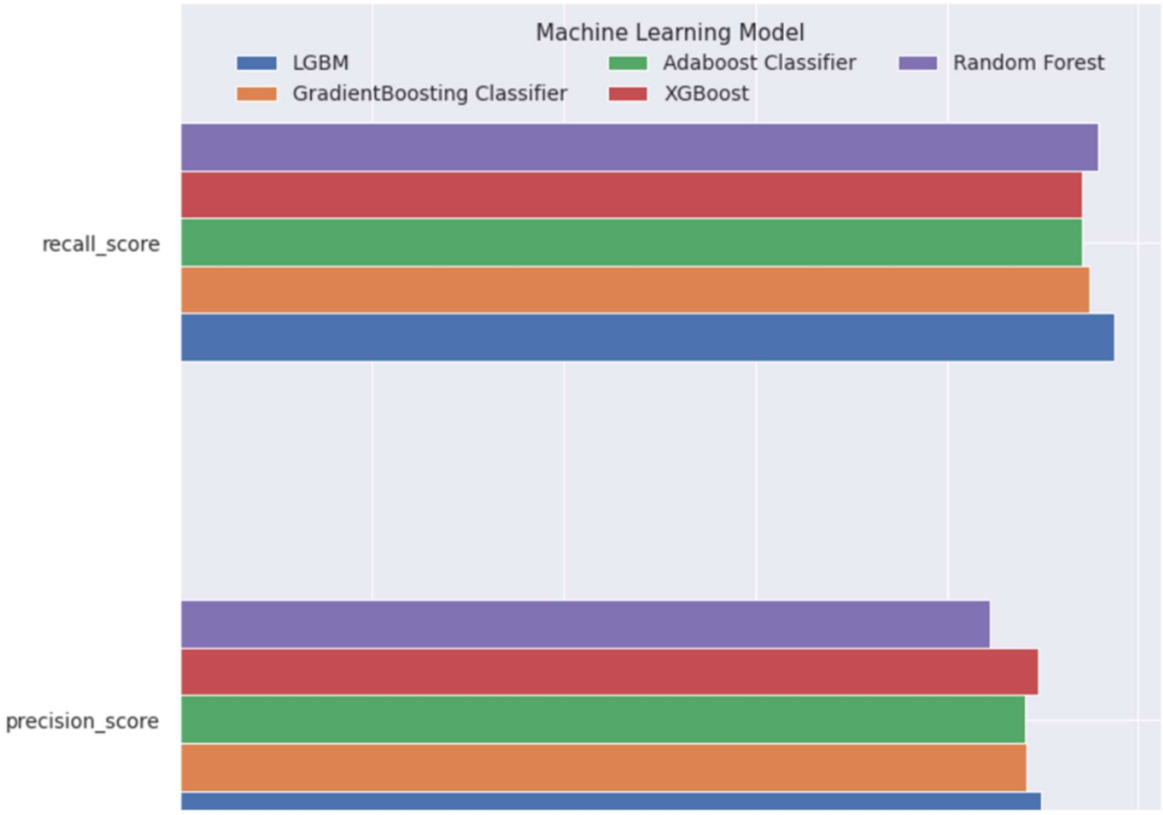
Random Forest

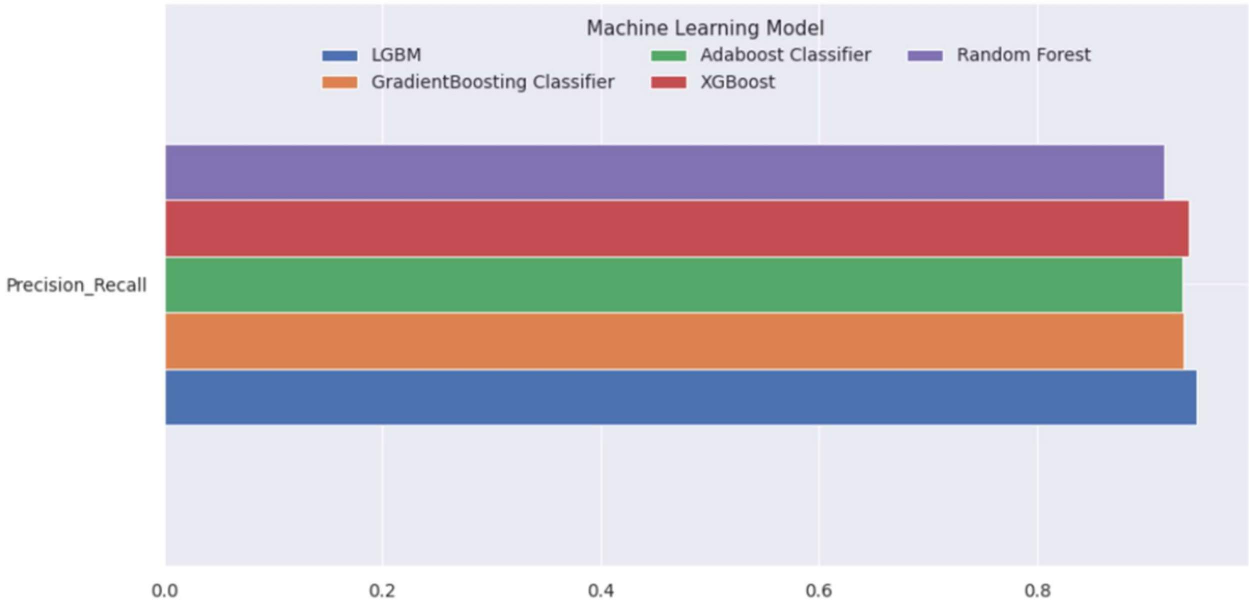
```python
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()

rf.fit(X_train,y_train)


y_pred_rf = rf.predict(X_test)

print("Training Accuracy :", rf.score(X_train, y_train))
print("Testing Accuracy :", rf.score(X_test, y_test))

cm = confusion_matrix(y_test, y_pred_rf)
plt.rcParams['figure.figsize'] = (3, 3)
sns.heatmap(cm, annot = True, cmap = 'YlGnBu', fmt = '.8g')
plt.show()

cr = classification_report(y_test, y_pred_rf)
print(cr)
print("-----------------------------------------")

false_positive_rate, true_positive_rate, thresholds = roc_curve(y_test,y_pred_rf)
roc_auc = auc(false_positive_rate, true_positive_rate)
print("ROC Curves              =",roc_auc)

precision, recall, thresholds = precision_recall_curve(y_test, y_pred_rf)
```

```
Training Accuracy : 1.0
Testing Accuracy : 0.8452380952380952
```

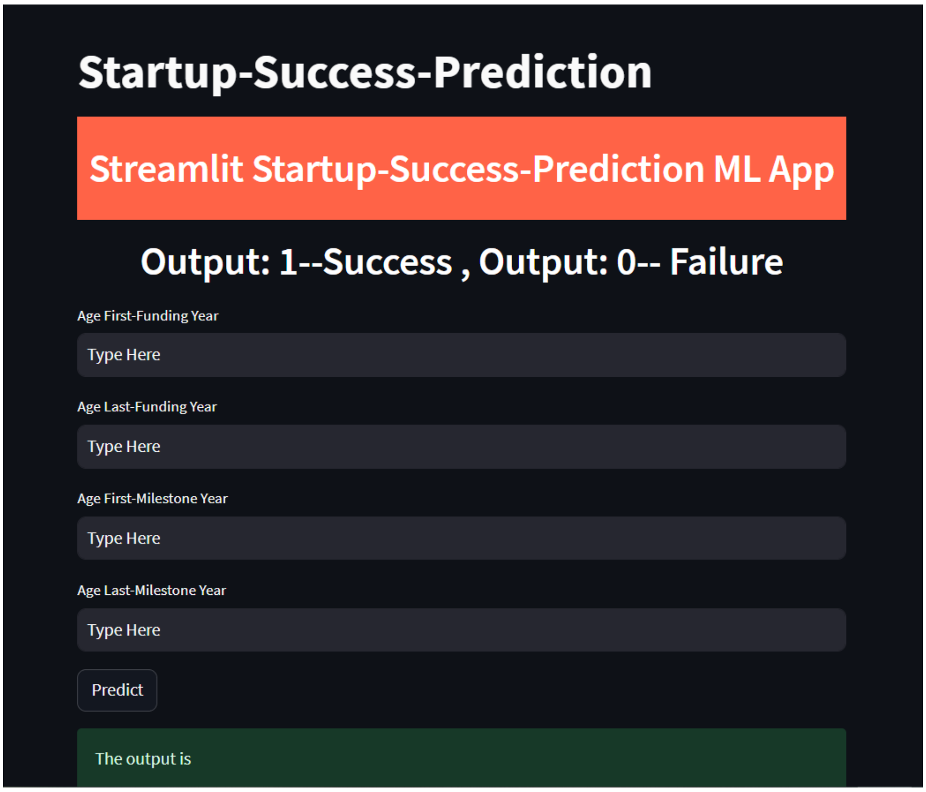


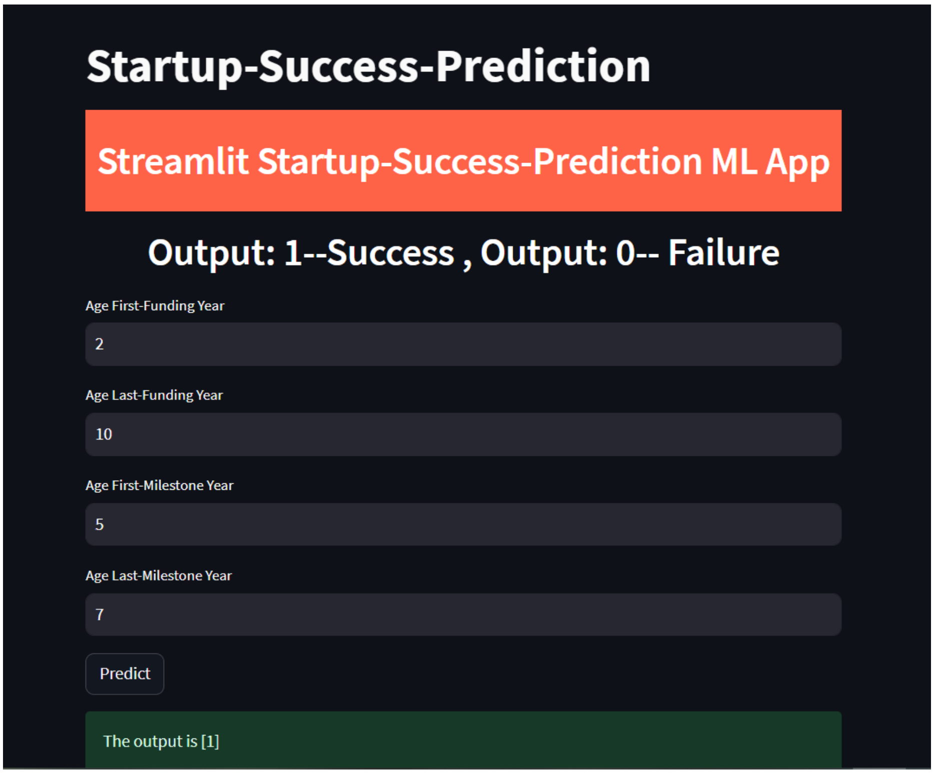

```
              precision    recall  f1-score   support

           0       0.85      0.57      0.68        49
           1       0.84      0.96      0.90       119

    accuracy                           0.85       168
   macro avg       0.85      0.76      0.79       168
weighted avg       0.85      0.85      0.84       168
```


Machine Learning Model
LGBM
GradientBoosting Classifier
Adaboost Classifier
XGBoost
Random Forest
recall_score
precision_score


Machine Learning Model
LGBM
GradientBoosting Classifier
Adaboost Classifier
XGBoost
Random Forest
Precision_Recall
0.0
0.2
0.4
0.6
0.8

Sample Output of Prediction App





Conclusion

In conclusion, the "Startup Success Prediction" project has been a comprehensive endeavor aimed at developing a machine learning model to accurately forecast the success of startups. Through various stages, including market and customer need assessment, external research, and dataset analysis, valuable insights have been gathered to understand the demand and requirements of such a solution. The prototype, designed with a user-friendly interface and powered by Python and machine learning libraries, demonstrates the project's commitment to providing a reliable and efficient platform for predicting startup success.

The proposed solution holds great potential in benefiting investors, venture capitalists, and other stakeholders in the startup ecosystem by assisting them in making informed decisions and optimizing their investment strategies. By accurately predicting startup success, the model can contribute to the growth and development of the startup ecosystem, fostering innovation and economic progress.

To maximize the project's impact, further validation and refinement are recommended. This could involve conducting extensive testing, gathering user feedback, and collaborating with industry experts to enhance the model's accuracy and reliability. Additionally, exploring suitable monetization strategies aligned with customer needs and market demand will be crucial for the project's sustainability and long-term success.

Overall, the "Startup Success Prediction" project has made significant strides in developing a valuable tool that leverages machine learning techniques to predict startup success. With its potential to empower decision-makers and shape the future of the startup ecosystem, the project stands as a testament to the power of data-driven insights and innovation.

References

- Link for project code:

https://colab.research.google.com/drive/1WpjnbQ8D6Og4hqDmhciToeF5gxIPDrcA?usp=sharing

- Link for dataset used:

https://www.kaggle.com/datasets/manishkc06/startup-success-prediction

- Link for the complete project website:

https://github.com/Rohit-925/startup_success_predictor_web_app