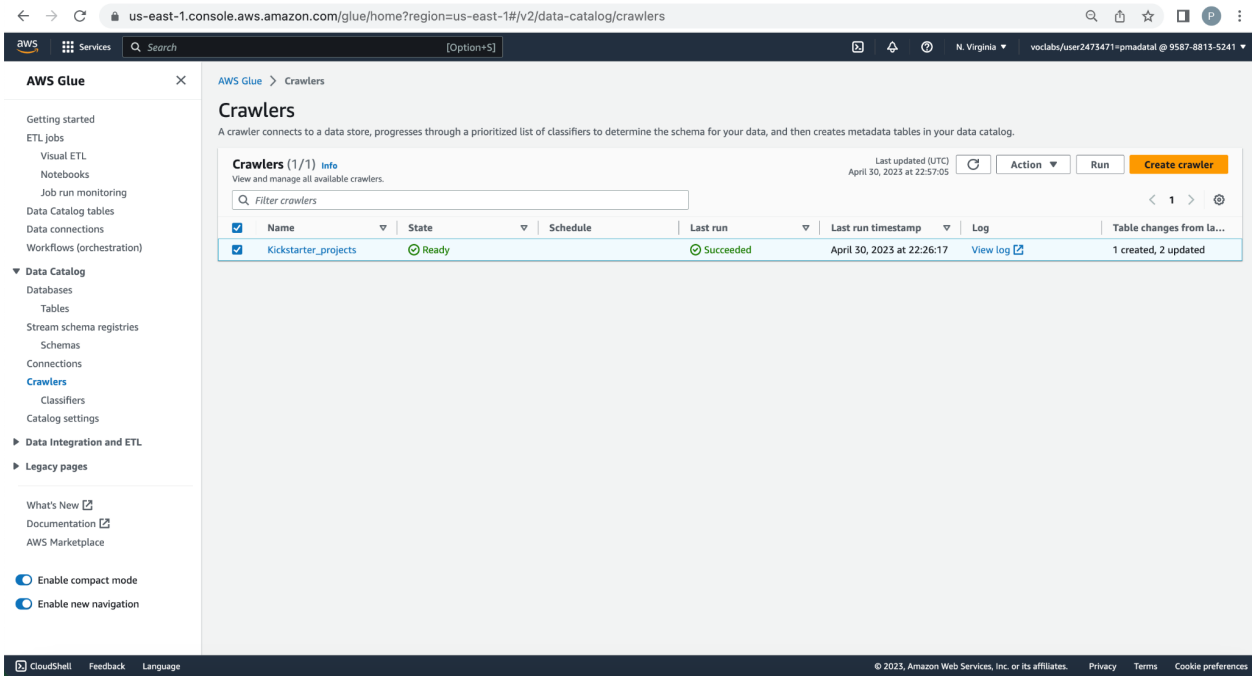
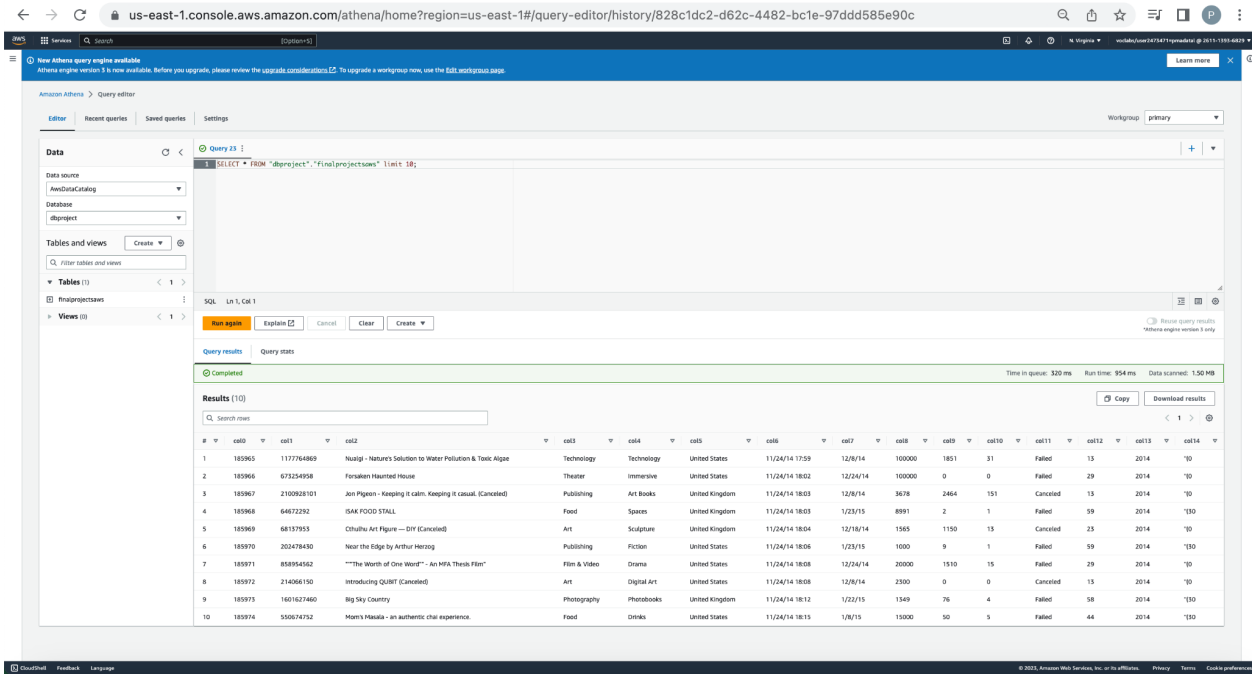
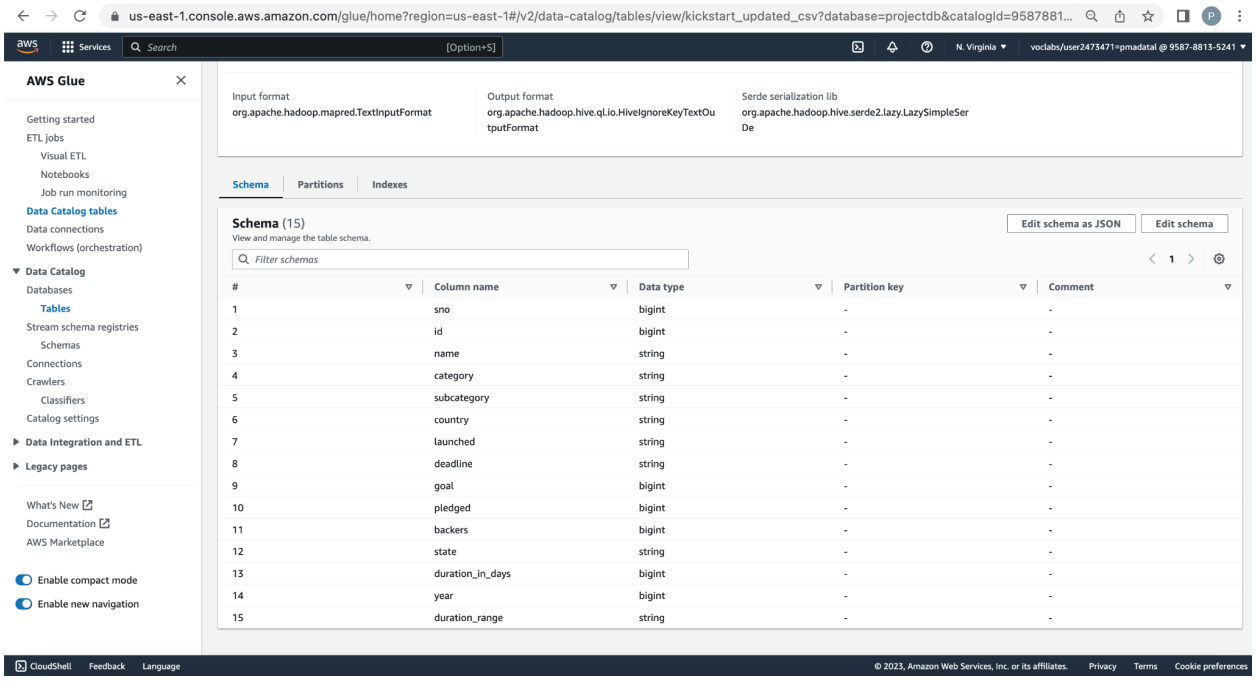
Analytics, Machine Learning:

We have used Amazon SageMaker, Athena, and Glue, which are great tools for recognizing and understanding the structure of the data. Athena allows us to query the data in S3 using SQL, while Glue automatically crawls and catalogs the data and generates ETL code to transform it into a format that is suitable for analysis. By using these tools, we have gathered a better understanding of our data and ensured that we are using the best ML tools for our specific dataset.

Additionally, Amazon machine learning technologies like SageMaker, Athena, and Glue can help extract insights from the data and build accurate machine learning models that can drive better business decisions.




us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/828c1dc2-d62c-4482-bc1e-97ddd585e90c
Amazon Athena > Query editor
Editor
Recent queries
Saved queries
Settings
Workgroup primary
Data
Data source
AwsDataCatalog
Database
dbproject
Tables and views
Create
Tables (1)
finalprojectsaws
Views (0)
Query 23
1 SELECT * FROM "dbproject"."finalprojectsaws" limit 10;
SQL Ln 1, Col 1
Run again
Explain
Cancel
Clear
Create
Query results
Query stats
Completed
Time in queue: 320 ms
Run time: 954 ms
Data scanned: 1.50 MB
Results (10)
Copy
Download results
Search rows


us-east-1.console.aws.amazon.com/glue/home?region=us-east-1#/v2/data-catalog/tables/view/kickstart_updated_csv?database=projectdb&catalogId=9587881...
Services
Search
[Option+S]
N. Virginia
voclabs/user2473471=pmadatal @ 9587-8813-5241
AWS Glue
Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)
Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings
Data Integration and ETL
Legacy pages
What's New
Documentation
AWS Marketplace
Enable compact mode
Enable new navigation
Input format
org.apache.hadoop.mapred.TextInputFormat
Output format
org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib
org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Schema
Partitions
Indexes
Schema (15)
View and manage the table schema.
Edit schema as JSON
Edit schema
Filter schemas
| Column name | Data type | Partition key | Comment
1 | sno | bigint | - | -
2 | id | bigint | - | -
3 | name | string | - | -
4 | category | string | - | -
5 | subcategory | string | - | -
6 | country | string | - | -
7 | launched | string | - | -
8 | deadline | string | - | -
9 | goal | bigint | - | -
10 | pledged | bigint | - | -
11 | backers | bigint | - | -
12 | state | string | - | -
13 | duration_in_days | bigint | - | -
14 | year | bigint | - | -
15 | duration_range | string | - | -
CloudShell
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Evaluation and Optimization:

AWS offers a variety of machine learning models and analytic tools, and we have used logistic regression and random forest models.
The likelihood that a dependent variable that is binary will take a specific value is predicted by a logistic regression model of the relationship between a binary dependent parameter and any number of independent variables.
Regression as well as classification are handled by an ensemble learning method known as random forest.It functions by building multiple decision trees on various subsets of the data that are randomly chosen, followed by combining the predictions from each tree to create a final prediction.

Results

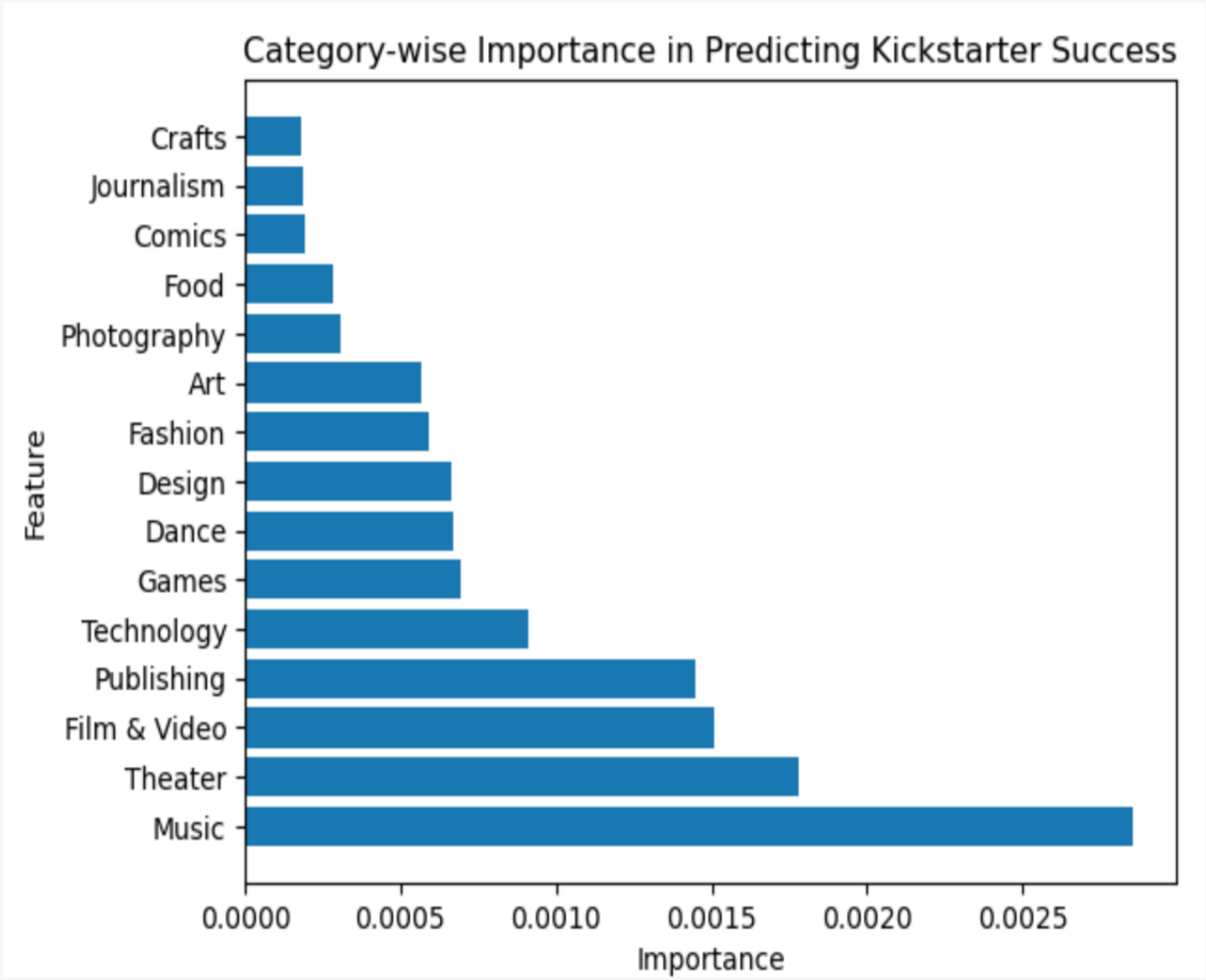
The accuracy of a machine learning model on the dataset can be a useful metric for evaluating the performance of the model. For our Kickstarter dataset, accuracy can tell you how well the model is able to predict the outcome of a crowdfunding campaign (i.e., whether it will be "successful," "failed," "canceled," or "suspended") based on the input features.

The notebook containing the code for building the models, evaluating their accuracy, and computing other relevant metrics can be accessed via the link provided below.

Link:
https://github.com/PavanKalyanReddyM/BDA_Project/blob/main/Deliverable_3.ipynb

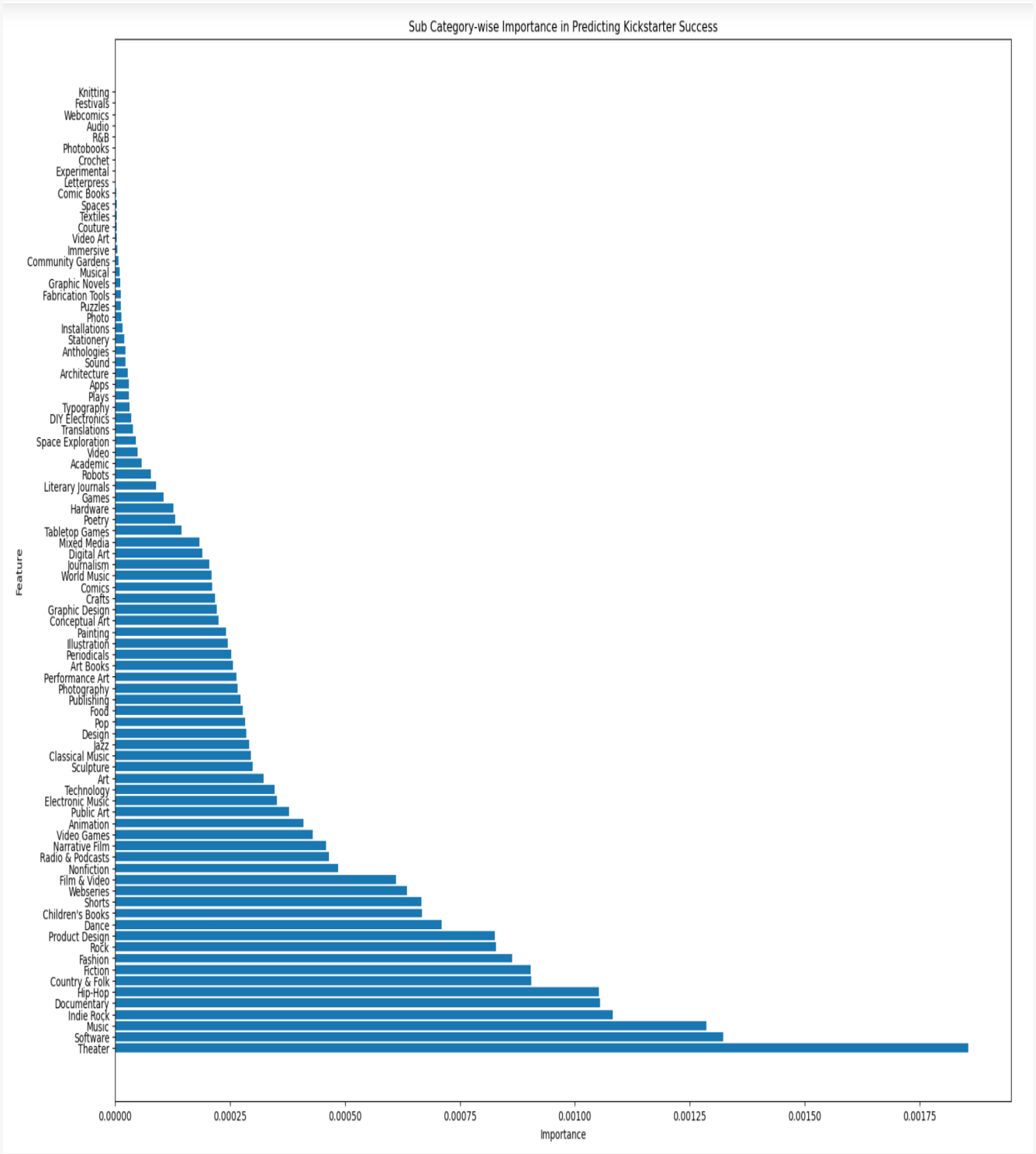
1. How important are categories in predicting the success of Kickstarter campaigns?

Feature Importance:

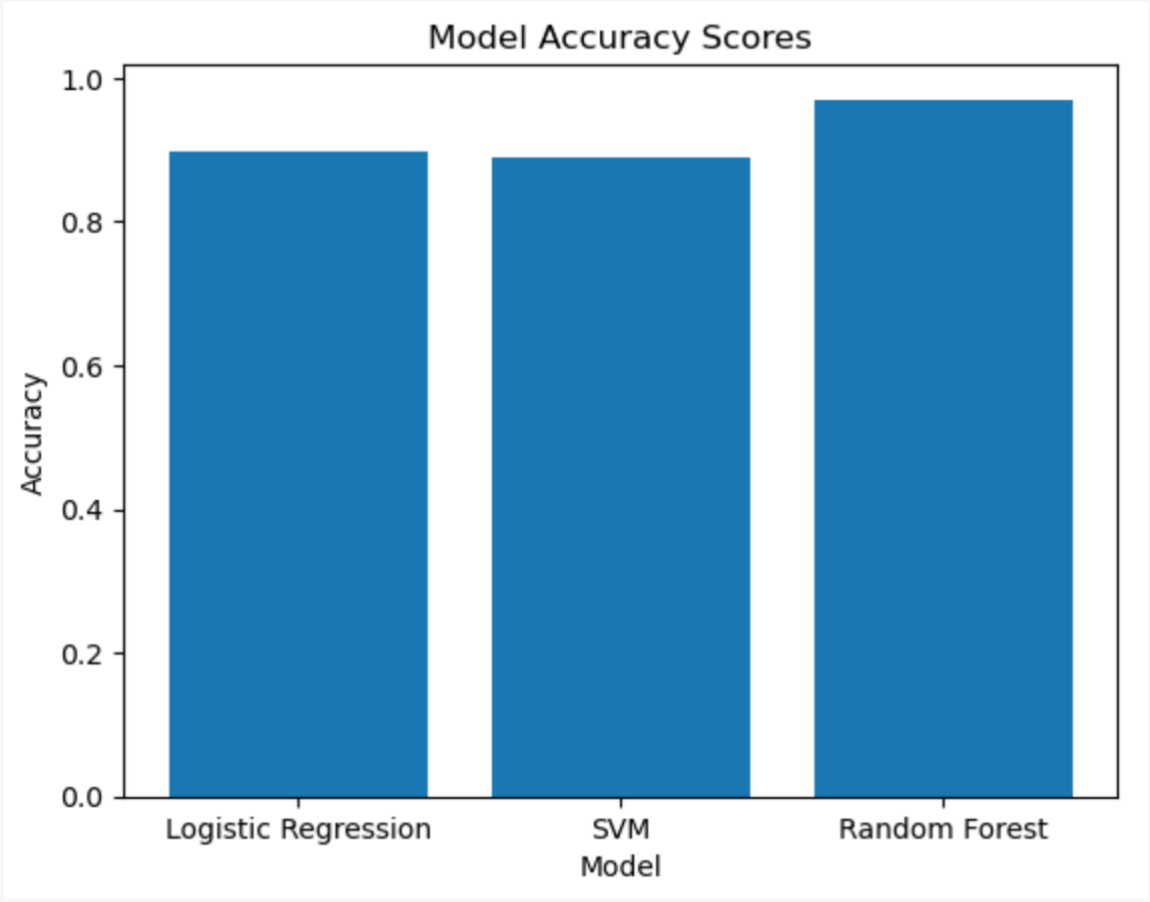


2. How important are subcategories in predicting the success of Kickstarter campaigns?

Feature Importance:



Model Accuracy Scores:



Based on the accuracy results, the Random Forest and Logistic regression models performed the best with an accuracy of 96% and 89%, respectively. SVM had an accuracy of 88%.

Considering the accuracy and other metrics, it can be suggested that the Random Forest model is the best option for predicting project success based on the available features.

The feature importance analysis revealed that the "Category" and “Subcategory” features has the highest importance in predicting the success or failure of Kickstarter projects.

Analyzing feature importance for category and subcategory wise can provide valuable insights into the factors that influence the success or failure of crowdfunding campaigns and can help guide decision-making and improve predictions in the future.

Based on these results, it may be beneficial to invest in Kickstarter projects in certain categories that have a higher success rate.

Future Work, Comments:

1) **The dataset is unique** because it provides a large and diverse sample of Kickstarter campaigns across different categories, subcategories, and years. This makes it a valuable resource for studying the factors that contribute to the success or failure of crowdfunding campaigns.

   Dealing with imbalance is an important issue when analyzing Kickstarter project data, as there is a significant class imbalance between successful and failed projects.

   Data cleaning for Kickstarter projects typically involves removing duplicates, dropping irrelevant columns, and dealing with missing values.

   Outlier treatment was also necessary for certain variables, such as the funding goal or number of backers.

   Imputation was used to fill in missing values in certain cases, such as when some variables were missing for only a small proportion of the observations.

2) In the initial stage of the project, we utilized AWS Athena and AWS Glue to locate the data and created dashboards in AWS Quicksight. In the Analytics and Machine Learning modeling phase, we employed regression models, such as linear regression, to present the data results. AWS SageMaker was used to generate the outcomes. Adding new attributes such as year and duration range was helpful in extracting additional information from the data and improving the predictive power of the model.

3) The analysis of the Kickstarter project dataset involved the use of various AWS resources, such as AWS Glue, AWS Athena, and AWS Quicksight. These tools were utilized for data identification and for a better understanding of the data after pre-processing. The processed data was then transferred to AWS SageMaker for the application of Machine learning models to obtain the final results.

4) During the initial stages of the project's creation, a few challenges were faced while working with the Kickstarter dataset. The pre-processing involved data cleaning and feature engineering to enhance the predictive power of the model. The dataset was sourced from Kaggle, and it was challenging to set it up on the AWS platform due to compatibility issues, especially generating schema in Glue after running the crawler in Athena. In conclusion, while the project presented several challenges, we overcame them by working collaboratively and leveraging the available resources to deliver a high-quality analysis of the Kickstarter dataset.

5) For future work, there are several other features, like project description, creator's background, social trends, etc., that can be included to improve the model's accuracy. Analyzing the success rate of campaigns across different states or cities can provide useful insights for creators looking to launch campaigns in specific locations.

Additionally, if the dataset is expanded to include other countries, regional analysis could be conducted.

6) If individuals want to use our work for the Kickstarter project dataset, which is available from Kaggle, they should set up an AWS account with appropriate credentials. The next step is to load the dataset into AWS and set up AWS Glue and AWS Athena to preprocess the data. Once the data is preprocessed, they can use AWS SageMaker to run the Jupyter notebook consisting of the code. The code consists of data cleaning, feature engineering, and model building. They can modify the code according to their requirements and run the notebook to obtain the results. All the necessary requirements for the procedure are mentioned in the Github repository, and individuals can refer to it for any further assistance.