Estimating Rocket Launch Costs With Data

Theophilus Ankrah

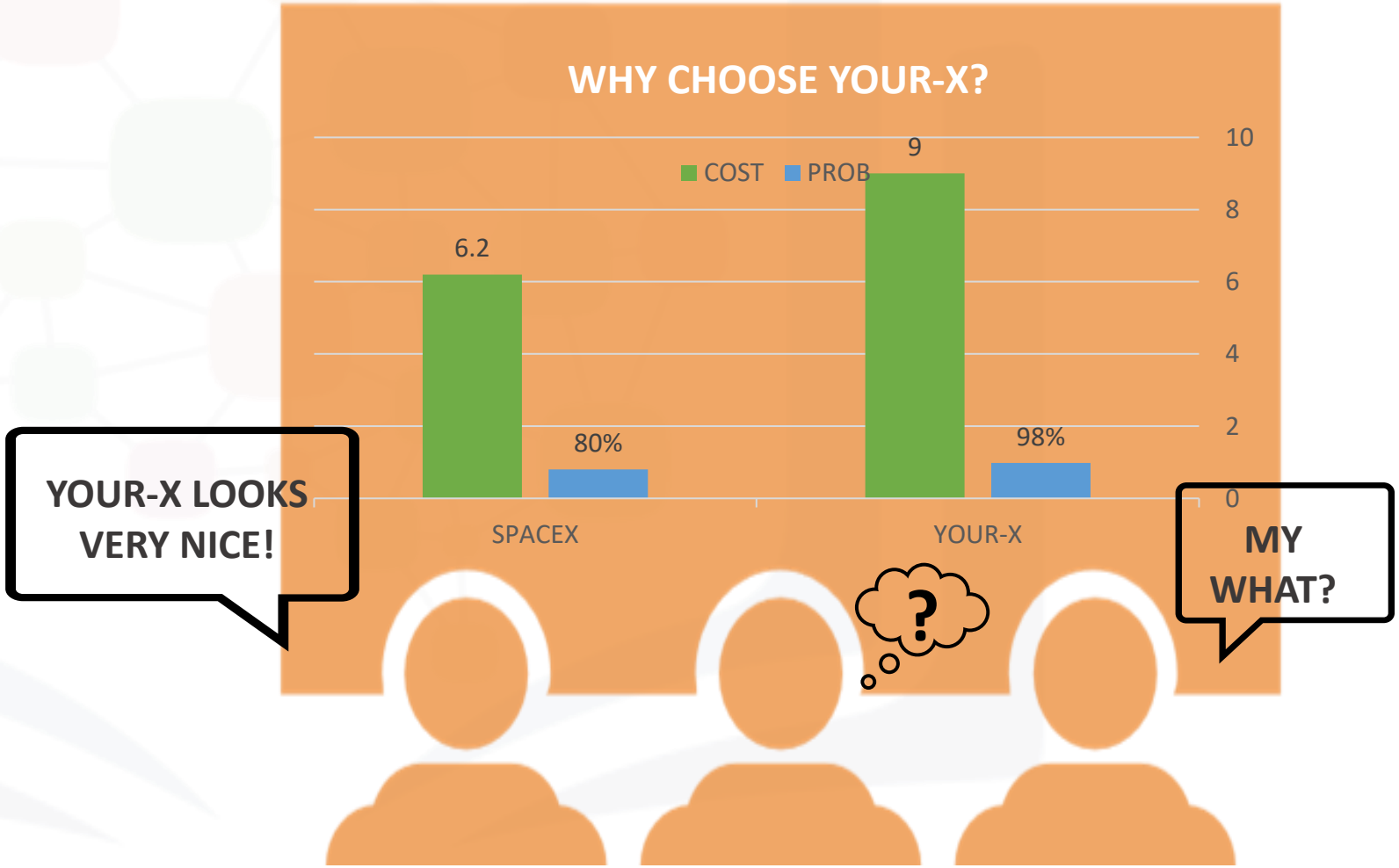
15th June 2023

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
- Conclusion



EXECUTIVE SUMMARY

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

INTRODUCTION

- Project background and context
  - Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.
- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.


METHODOLOGY

METHODOLOGY

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Classification models were defined and tuned to obtain the best parameters

Data Collection

The data was collected using various methods

- SpaceX API was selected as the primary data source – Achieved using get requests.
- Data was collected converted in a pandas dataframe. – Achieved by normalizing json formatting
- Collected data was cleaned, checked for missing values and replaced them. - Using pandas
- Falcon 9 launch records were extracted from Wikipaedia. – Achieved through web scrapping
- The collected data was converted into a tabular dataframe for further analysis. – Using pandas

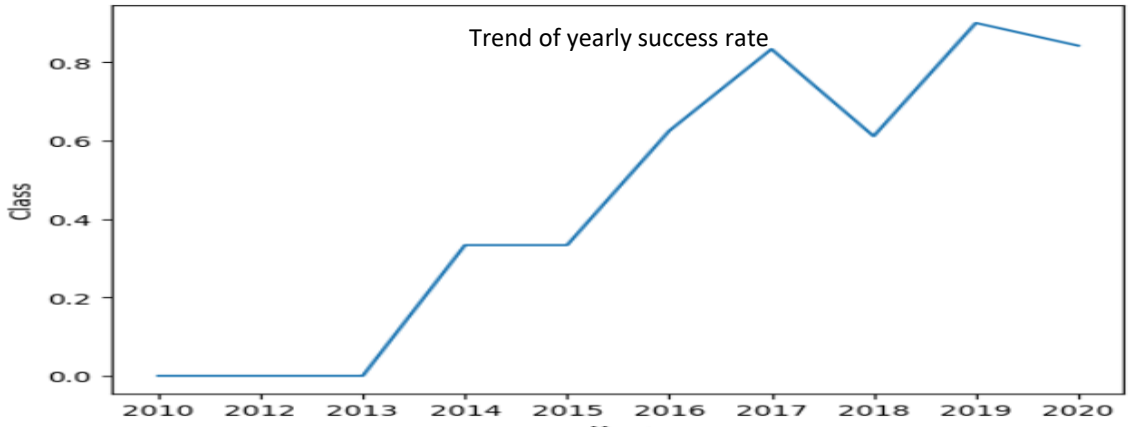
Data Wrangling

- Exploratory data analysis was initiated to determine training labels for our models.
- Calculated columns for the totals for all Falcon 9 launch sites and target orbits were found and included for further analysis.
- Landing outcomes were categorized based on retrieved data and placed in a target column for our model.

EXPLORATORY DATA ANALYSIS VIA DATA VISUALIZATION

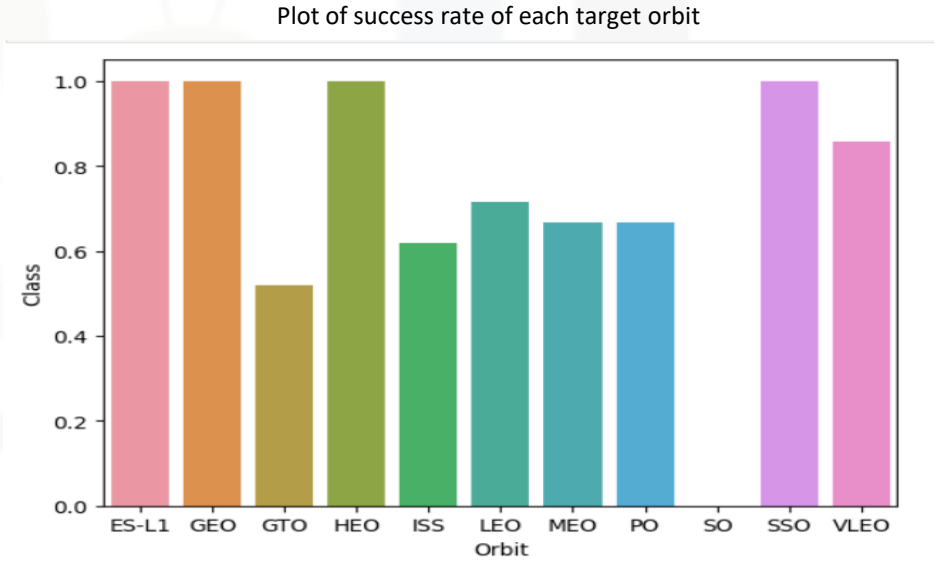
Using data visualizing the following relationships were discovered between:

- flight numbers and launch sites
- payloads and launch site,
- success rate of each orbit type,
- flight number and orbit type,



Key take aways :

1. An upward trend was observed for yearly launch success trend.
2. Target orbits had varying success rates.



EXPLORATORY DATA ANALYSIS VIA SQL

Further exploratory analysis was carried out using SQL via an IBM DB2 connection

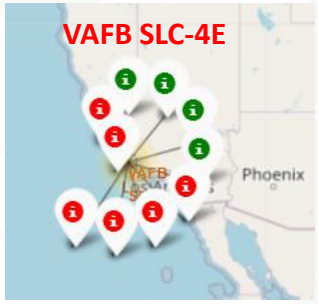
Key insights:

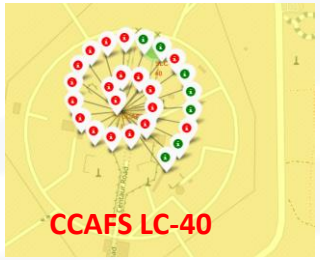
1. Average payload for Falcon 9 v1.1 = 2,928.4 kg
2. Total Successful missions for Falcon 9 = 100
3. Total Failed missions for Falcon 9 = 1

Building Interactive Maps with Folium

- All launch sites were marked on the map – Observation made for all launch site being close to the beach and sea.
- Successful launches were marked with green markers with a class assignment of 1 and 0 & red markers for unsuccessful launches. – KSC LC – 39A had the highest success rate
- Site success rates were emphasized through marker clusters that clearly tell the general success rate for each site.
- Nearness to key landmarks was also investigated for possible relevance.

Success rates by Launch Sites via marker clusters

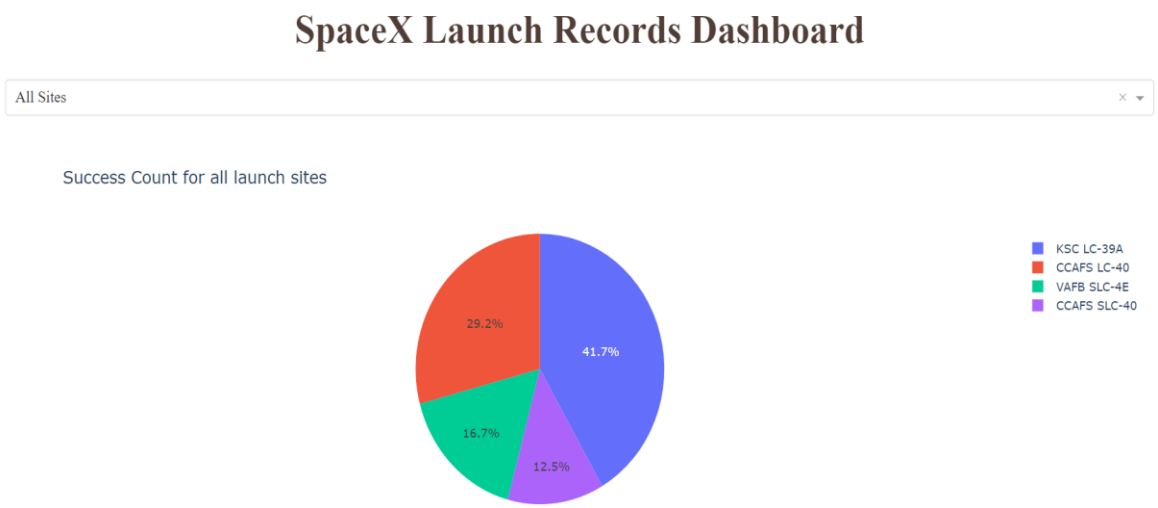




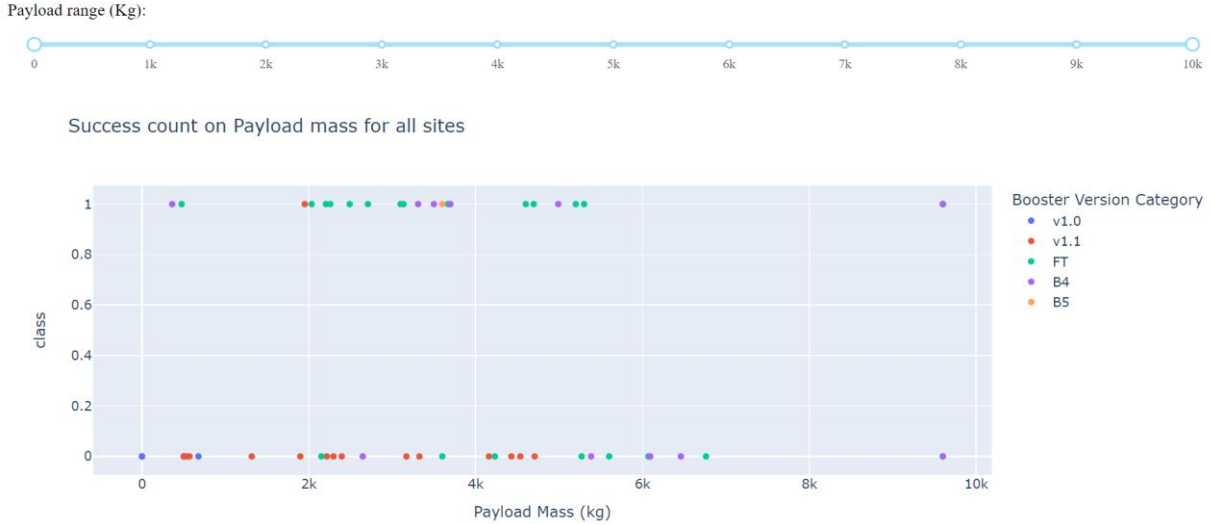
PLOTLY DASHBOARD

An interactive dashboard was developed with the following properties:

1. A pie chart to communicate successful launches by certain sites.



2. A scatter plot to show the relationship between Mission Outcome and Payload Mass for the different booster versions



Predictive Analysis

- The objective of this analysis was to produce a predictive model that can predict with a high accuracy whether a launch will be successful or not.
- The collected data is transformed and normalized before being split into training and testing data – 80% used for training and 20% for testing.
- Several models were developed, and their parameters tuned to obtain the highest performing model parameters.
- The prediction accuracy of all the models was compared and found as:

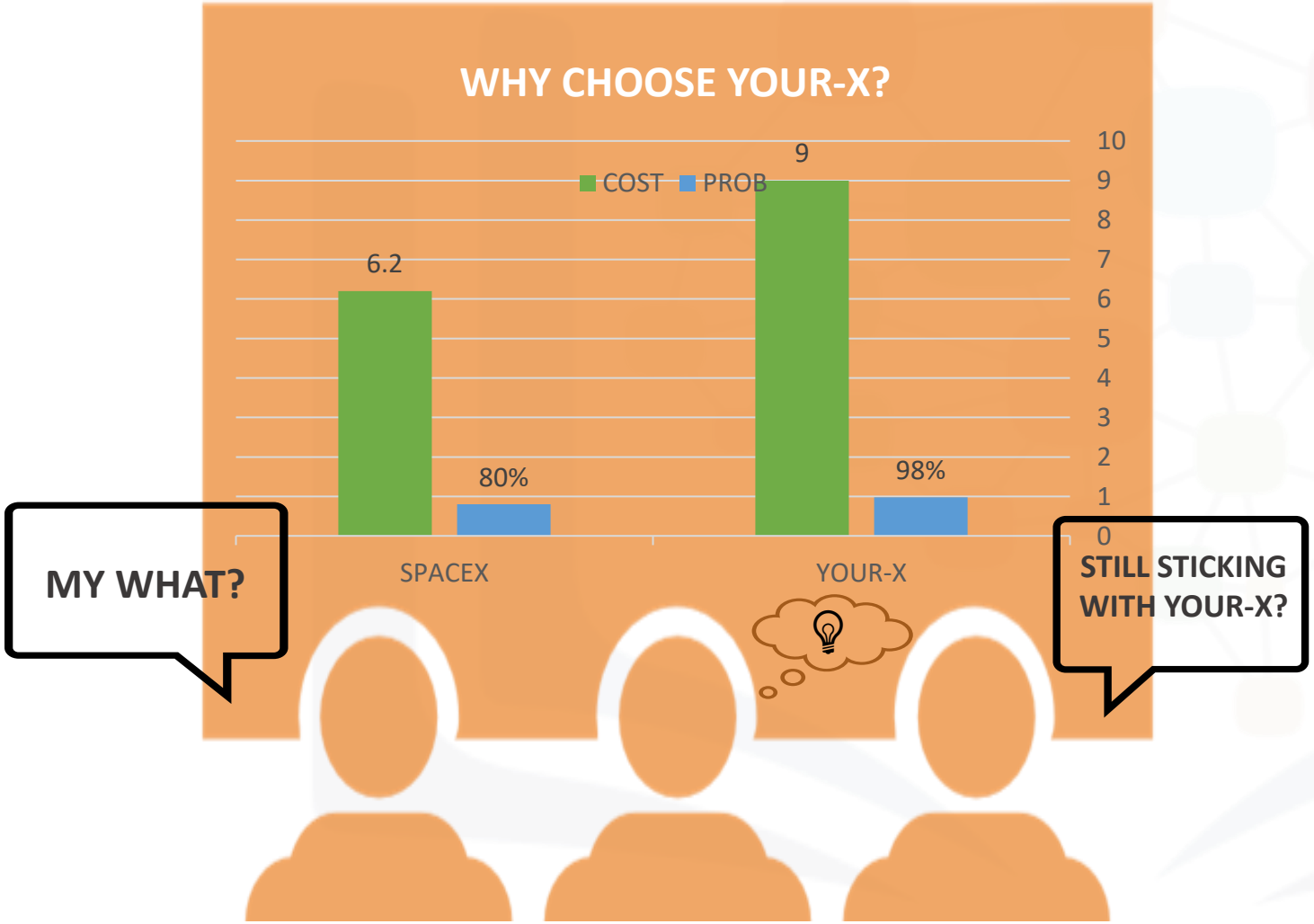
155]:

	Model	Accuracy Score
0	LOGISTIC REGRESSION	0.833333
1	SUPPORT VECTOR MACHINE	0.833333
2	DECISION TREE	0.888889
3	K-NEAREST NEIGHBOR	0.833333

Key insights:

1. Almost all models had similar accuracy scores
2. In this case, Decision tree fared best after multiple runs
3. However, Decision tree accuracy changes at every run and is sometimes the lowest
4. All models turned out high accuracies

DISCUSSION



- Was enough data collected for the project?
- Was data credible and unbiased?
- Did the methodology lack in any way?
- Were the dependent variables enough to inform the model?
- Were the right models selected?
- Did the tuning improve model delivery?
- Can the model be improved?
- Was the objective of this project achieved?

CONCLUSION

We can conclude that:

- SPACEX gets better success rates with increasing number of launches
- KSC LC-39A had the most successful launches of any sites.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- Launch success rate started to increase in 2013 till 2020.
- The Decision tree classifier is not the best machine learning algorithm for this task as with each run , the accuracies keep changing.
- Logistic regression model is recommended as we can also check for the probability of a prediction and has equally high accuracy as the others.