Final Module Project: Data Science Capstone

Marco Mudenge

September 6^{th} 2023

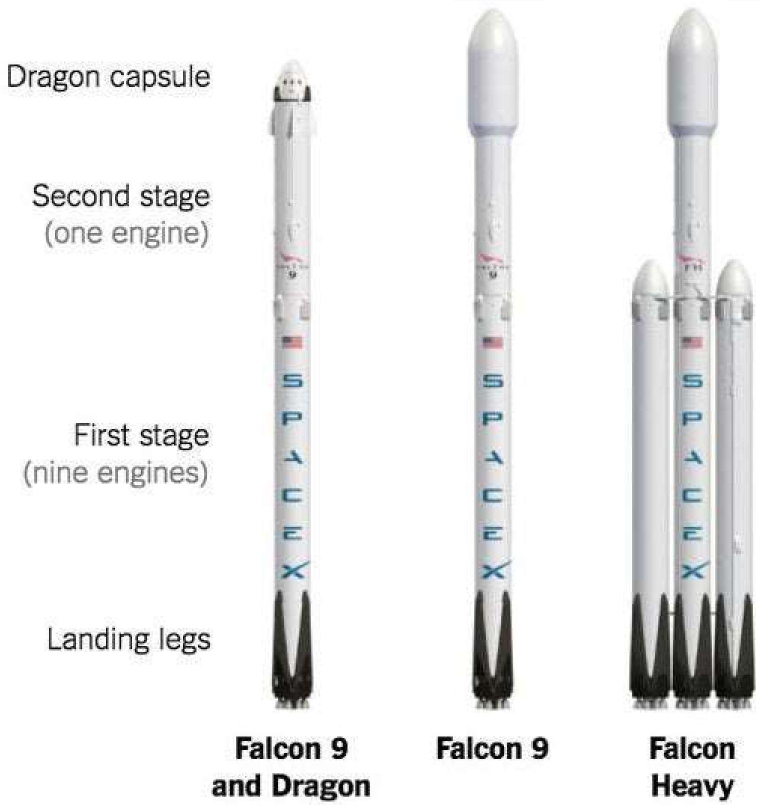
IBM Developer

SKILLS NETWORK

EXECUTIVE SUMMARY

- Determining the landing success of SpaceX Falcon 9 rocket has a importance for competitors biding
- SpaceX provides data sets that allow us to investigate the problem
- Exploratory Data Analysis, Data Visualization and Predictive Analysis helps us better determine the landing success

INTRODUCTION



- SpaceX advertises the Falcon 9 rocket at 62 millions
- The Falcon 9 cost way less than competition because its first stage can be reused
- Determining if the first stage will land is of great importance

- Goal of this project: predict the Falcon 9 landing success according to data

METHODOLOGY: DATA COLLECTION & WRANGLING

METHODS

- Collect base data using the SpaceX API
- Replace key IDs with info using the API
- Filter data, keep only Falcon 9 launches
- Replace missing values with mean

GOALS

- Retrieve the relevant SpaceX data
- Reorganize the data

OUTCOME

- **Data set #1** created for data wrangling

METHODOLOGY: EDA - Wrangling (Part 1)

METHODS

- Using data set #1
- **Explore** occurrences of landing site and **orbits**
- **Explore** occurrences of different **mission outcomes**
- **Create** a binary **column** for the landing outcome **success**

GOALS

- Find patterns in the data
- Determine a label for our predictive supervised models

OUTCOME

- **Data set #2** created for predictive analysis to be used as labels

METHODOLOGY: EDA with Visuals (Part 2)

METHODS

- **Using data set #2**
- Explore **relations between variables** using scatter plots & bar charts
- Explore launch success **yearly trend** with line plot
- **Features engineering** : Use One-Hot encoding to create new numeric features

GOALS

- Find patterns in the data
- Perform data feature engineering

OUTCOME

- **Data set #3** created for predictive analysis used for the features

METHODOLOGY: PREDICTIVE ANALYSIS

METHODS

- Using data set #2 (label) & #3 (features)
- **Standardize** the features
- **Split the data** into training/test sets
- Train & test **supervised models** :
  - Logistic regression, SVM, Decision trees & KNN
- Use a **Grid Search** to find the best parameters
- **Evaluate** the models

GOALS

- Find best hyperparameters for each models
- Determine the best models

OUTCOME

- **Results** for each predictive models

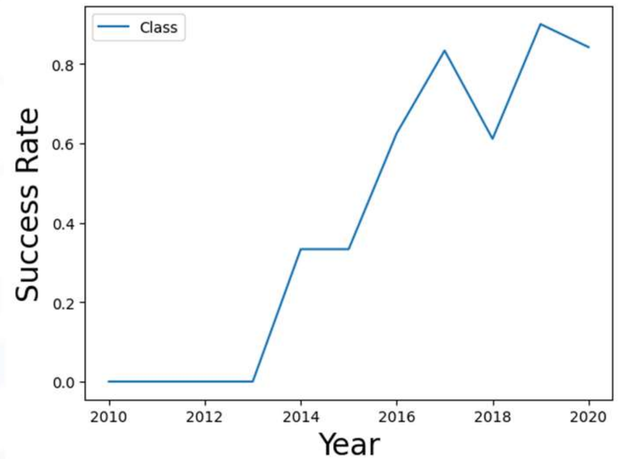
METHODOLOGY: VISUAL ANALYTICS

- Point1
- Point2
- Point3
- Point4
    - Sub Point1
    - Sub Point2

RESULTS: EDA with Visualisations

Variables	Plot	Findings
FlightNumber, Payload, LaunchSite, Orbit (Multiple combinations)	Scatter	No patterns
Orbit vs Class	Bar	No patterns
Year vs Class	Line	Growing success rate since 2013

- No conclusive patterns detected
- Possibles causes : attributes like FlightNumber & LaunchSite are not values with empirical meaning.
- Growing success rate may be attributed to multiple external reasons such as growing expertise.



RESULTS: EDA with SQL

Better understanding of our dataset

Knowledge of distincts values and counts of:

- Landing outcome
- Mission outcome
- Launch site
- Booster versions
- Sum and average payloads

Mission_Outcome	COUNT(MISSION_OUTCOME)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

RESULTS: Interactive Map

Features of the interactive map:

- Each launch sites marked on the map

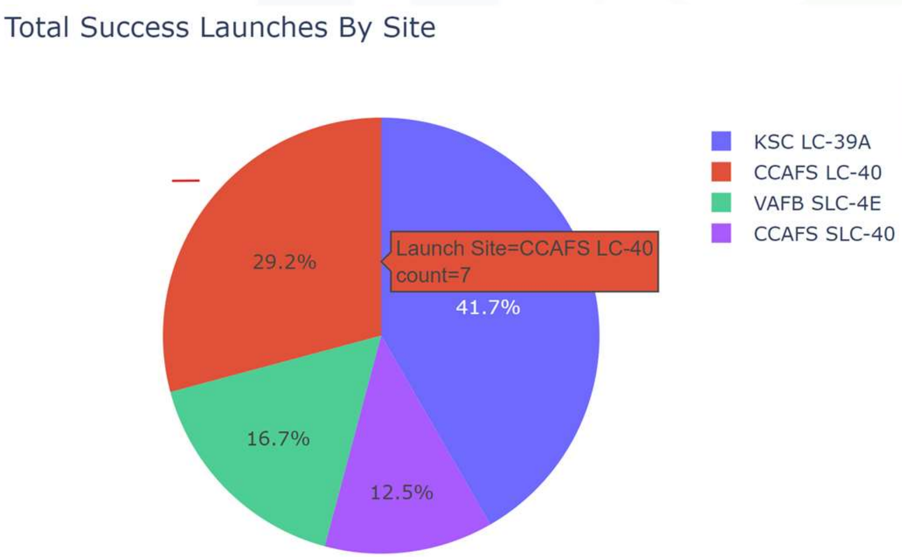
- Clear indication of success/failure of each launch (green/red)

- Annotation of distance between a launch site and the nearest coastline

RESULTS: Plotly Dashboard

The pie chart shows a majority of successful launches are within 2 sites

We can not observe a pattern when plotting the payload against launch success



IBM Developer

SKILLS NETWORK

RESULTS: PREDICTIVE ANALYSIS

BEST PARAMETERS

- Logistic Regression
  'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'
- SVM
  'C': 1.0, 'gamma': 0.03162277660168379, 'kernel': 'sigmoid'
- Decision Trees
  'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 10, 'splitter': 'random'
- KNN
  {'algorithm': 'auto', 'n_neighbors': 4, 'p': 1}

MODEL RESULTS

Algorithm	Jaccard	F1-score
Logistic Regression	0.666667	0.800000
SVM	0.714286	0.833333
Decision Tree	0.769231	0.869565
KNN	0.714286	0.833333

DISCUSSION

- While no clear patterns were detected through EDA, supervised method were still able to predict landing outcome with up to 80% of accuracy.
- This lab provided an overview of different tasks of Data Sciences (Collection, Wrangling, Analysis, Visualization & Machine Learning).

CONCLUSION

- Data collection & Wrangling prepared SpaceX data
- EDA through visualisations and SQL provided a better understand of our data
- No clear patterns observed through the EDA
- Interactive map shows the location of landing sites
- Plotly Dash app was created to complete analysis
- Supervised models are around 80% accurate in predicting the outcome of landing