

**Introduction**

The Applied Data Science Capstone Project was a pivotal component of my journey to achieve the IBM Professional Certification in Data Science through Coursera. This project was an opportunity to apply the skills I acquired over a seven-month period, utilizing real-world datasets for extraction, cleaning, exploration, and analysis to derive meaningful insights. I'm delighted to announce the completion of my certification and eager to share my experiences with the capstone project.

**Problem Statement**

SpaceX, a trailblazing company in aerospace manufacturing, space transportation services, and communications, shares its disruptive nature with its sister company Tesla, both founded by Elon Musk. Despite its relatively young age of less than two decades, SpaceX has significantly reduced launch costs by over 50% compared to its competitors and aims for a 99% reduction upon the completion of the Starship project.

This cost-efficiency is largely due to SpaceX's innovative technology, allowing the first stage booster, which accounts for 70% of the rocket's cost, to be landed and reused. This breakthrough in reusable technology has enabled SpaceX to dominate the market, with reused boosters costing 50% less than new ones.

In this capstone, we will analyze data extracted from Wikipedia and the SpaceX API through web scraping. Our goal is to gain insights and predict the successful landing of boosters on drone ships.

**Data Collection and Wrangling**

The initial step involves data collection, followed by an examination for missing data and assessment of data types. Our data cleaning process includes:

* Replacing missing data, possibly using the mean.
* Modifying data types.
* Representing categorical data with integer or float values through techniques like one-hot encoding.

**Exploratory Data Analysis**

Once the data is cleansed, we move onto exploratory analysis. This involves using visualization techniques to gain deeper insights into the launch data and patterns.

Here are a few selected screenshots :

A white background with orange and blue dots

Description automatically generated

From the graphical representation, it can be inferred that:

1. Initial space missions predominantly originated from the CCAFS-SLC-40 launch site, succeeded subsequently by launches from KSC-LC-39A.

2. The majority of space missions have been launched from CCAFS-SLC-40.

3. There have been a relatively smaller number of launches from the VAFB SLC 4E site.

A white background with orange and blue dots

Description automatically generated

Based on the visualization, the following conclusions can be drawn:

- VAFB SLC 4E is primarily associated with launches of low payload.

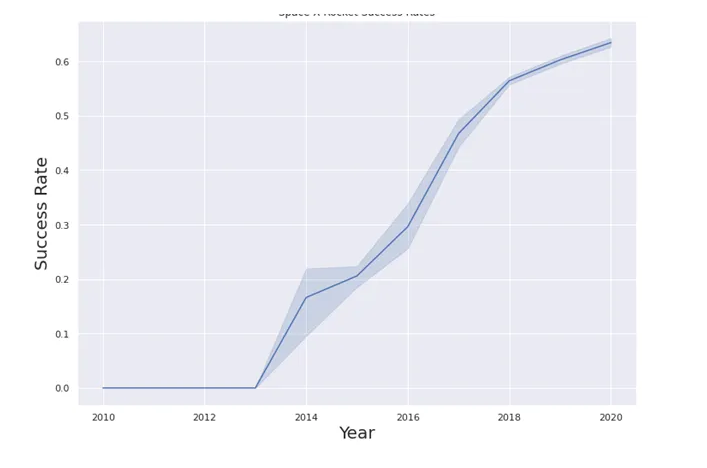
- CCAFS SLC 40 has a higher frequency of both high payload and low payload launches.

A graph of different colored bars

Description automatically generated

The visualization indicates that:

* GEO (Geostationary Orbit), HEO (Highly Elliptical Orbit), and ES-L1 (Earth-Sun Lagrange Point 1), along with SS (Sun-Synchronous) orbits, exhibit a high rate of successful missions.

 Upon examining the data, it is observed that there is a trend indicating an increase in the probability of successful landings.

**Predictive Analysis**

A chart of different colors

Description automatically generated

I utilized the data to train various machine learning models, including the KNeighborsClassifier.

A screenshot of a graph

Description automatically generated

Decision Tree classifier

A graph of different colored squares

Description automatically generated

Logistic regression

A chart of different colors

Description automatically generated

Support Vector Machine

A graph of different colored squares

Description automatically generated

svm confusion matrix

I refined the models, focusing on evaluating their scores, best scores, and confusion matrix plots to determine the most accurate model. It was concluded that the KNN (K-Nearest Neighbors) model outperformed the others in terms of score, accuracy, and demonstrated the least bias in its confusion matrix.

.

A screenshot of a computer program

Description automatically generated

A screenshot of a graph

Description automatically generated

Interactive Map Utilizing Folium

Given that SpaceX launches originate from various sites, I have illustrated the data of both failed and successful launches using clustered markers on the map. By zooming in and out, you can observe the distinct clusters representing successful and failed launches.

A map with orange circles and dots

Description automatically generated

Interactive Dashboard Creation Using Plotly Dash

Plotly Dash, a Python library, simplifies the process of developing dashboards for data scientists. This tool enables the creation of interactive dashboards where one can alter inputs to dynamically view the corresponding changes in graphical representations.

A colorful pie chart with numbers

Description automatically generated

A graph with colorful circles

Description automatically generated

From my analysis, I found that the majority of launches were conducted at the Kennedy Space Center. Reasons for this preference include its proximity to SpaceX's production facility.

Kennedy Space Center's Pad 39A was the primary launch site, particularly for missions to Very Low Earth Orbit (VLEO), Geostationary Orbit (GEO), or the International Space Station (ISS). This makes it an ideal location for such launches.

Falcon Heavy often carries full payloads, optimizing the capacity of the Falcon's payload.

Over time, the probability of successful booster landings has improved, benefiting from the data gathered from previous failures.

SpaceX achieved its first successful booster landing on May 6, 2016.

**Conclusion**

By utilizing existing data for analysis, SpaceX and other rocket companies can identify more efficient ways to reduce launch costs. This evolution is crucial to prevent their traditional, more expensive launch methods from becoming obsolete and losing their market share.