IBM DATA SCIENCE CAPSTONE PROJECT

BY
NEHA REDDY
PALNATI

OUTLINE

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

EXECUTIVE SUMMARY

Methodologies Summary

Data collection

Data Wrangling

EDA Data visualization.

EDA with Sql .

Predictive Analysis

Interactive Map With Folium

Interactive Dashboard Using Plotly

Results Summary

- **Exploratory Data Analysis**
- **Predictive Analysis results**

INTRODUCTION

- **Background :SpaceX a rocket company launches satellites at low price like 70% less than their competitor since they lan their satellites for reusing them to launch .**

- **Problem:We use the previous data of launches of Falcon 9 rocket to predict the probability of the booster landing back to the pad influenced/correlated with the space launch site,the payload orbit,mass,landing pad location and the version of the booster.**

METHODOLOGY

- Data Collection -API & Web Scraping.
- Data wrangling-Extracting Load & Transform .
- Cleaning data to values that we can use -example labels to dummy integers .
- EDA with visualization and Sql.
- Interactive with Folium and Plotly Dash .
- Predictive Analysis -using Machine Learning Models .

METHODOLOGY

DATA COLLECTION

REST API: Using the rest api we extract the data in form of JSON and transform it to a dataframe using inbuilt python pandas method normalize .

WEB SCRAPING : Web scraping spacex launches from wikipedia and converting it into a dataframe .

REST API

```
json_data=requests.get(static_json_url).json()
```

Make request

```
# Use json_normalize meethod to convert the json result t
data=pd.json_normalize(json_data)
```

Normalize to df

Filter Falcon 9 only

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9=df_launch[df_launch['BoosterVersion']!='Falcon 1']
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

Create a dictionary for creating a dataframe from the dataset collected

Save to CSV

```
data_falcon9.to_csv('dataset_part_1.csv',index=False)
```

Wl

```
# use requests.get() method with the provided static_url
# assign the response to a object
response=requests.get(static_url)
```

Get content of the wiki

Loop Through and add column names

```
column_names = []
temp = soup.find_all('th')
for x in range(len(temp)):
    try:

        name = extract_column_from_header(temp[x])
        if (name is not None and len(name) > 0):
            column_names.append(name)
    except:
        pass
```

Create a Dataframe for the important columns

Loop through the request content and extract data

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

Save to cvs

DATA WRANGLING

In the data set, there are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean. True RTLS means the mission outcome was successfully landed to a ground pad False RTLS means the mission outcome was unsuccessfully landed to a ground pad.True ASDS means the mission outcome was successfully landed on a drone ship False ASDS means the mission outcome was unsuccessfully landed on a drone ship.

EXPLORATORY DATAANALYSIS

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Save to CSV

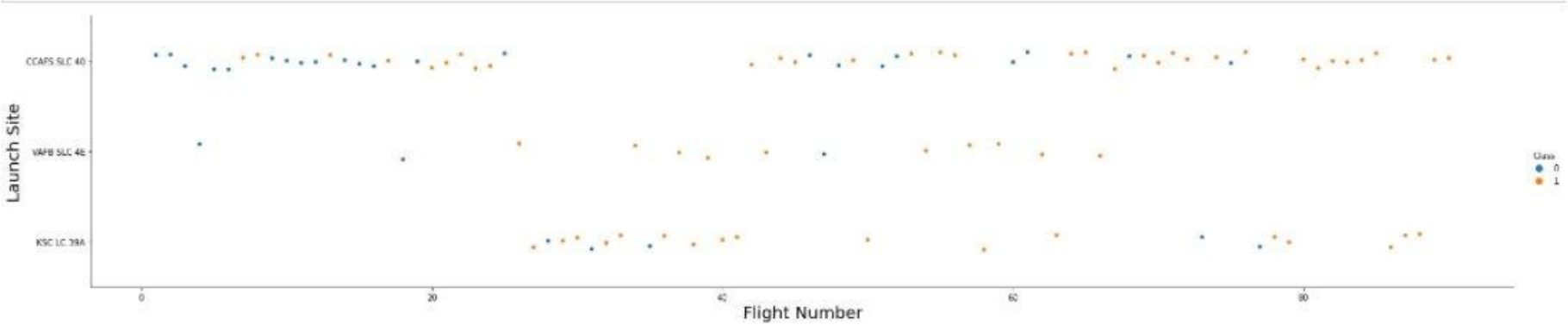
Calculate the % of Null Values

```
df.isnull().sum()/df.count()*100
```

EXPLORATORY DATA ANALYSIS WITH VISUALIZATION

- **Flight number & Launch Sites-Visualizing the launch from every site .**
- **Payload & Launch Sites-Payload launch from sites**
- **Success rate & Orbit type-Success rate compared to the orbit type**
- **Flight number & Orbit Type -Type of orbit for each launch**
- **Payload & Orbit type -Payload and the orbit .**
- **Trend of success rate-Trend of the success rate over the years .**

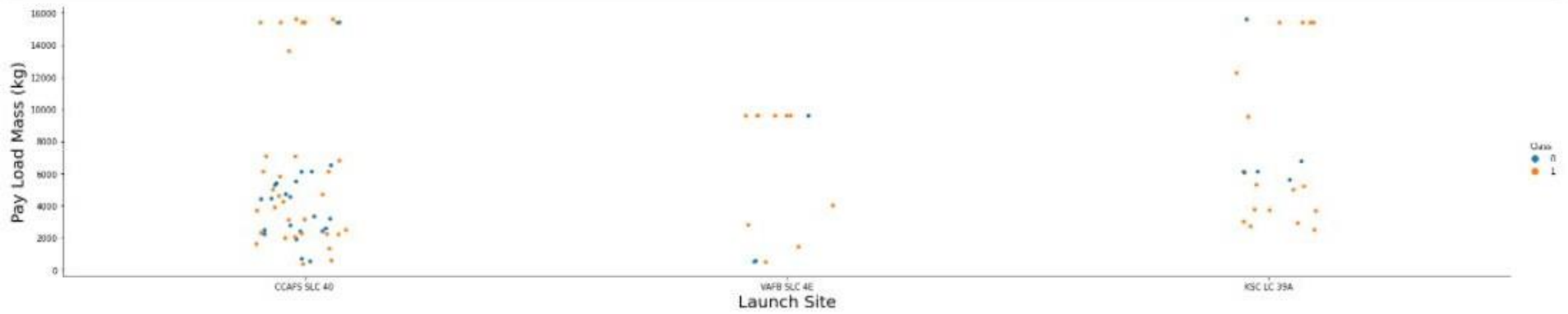
Flight number & Launch Sites



From the Visualization we can concluded that:

- Earlier flights lauch were from CCAFS-SLC-40 site ,Followed by KSC-LC-39A
- Most Launches are Launched from CCAFS-SLC-40
- Fewer Launches from **VAFB SLC 4E site**

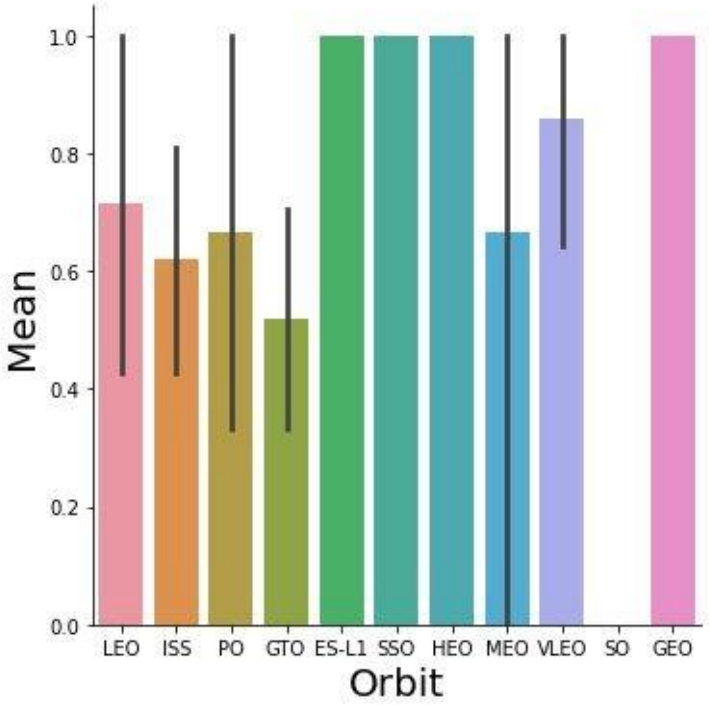
Payload & Launch Sites



From the Visualization we can concluded that:

- VAFB SLC 4E has Low Payload launches
- CCAFS SLC 40 has more Higher Payload Launches and Low Payload Lauches

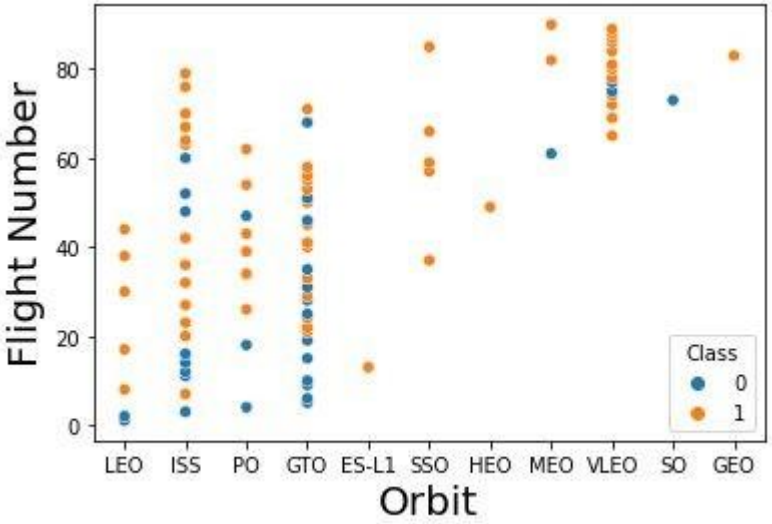
Orbit Success



From the Visualization we can concluded that:

- GEO,HEO & ES-L1,SS) have high success rate .

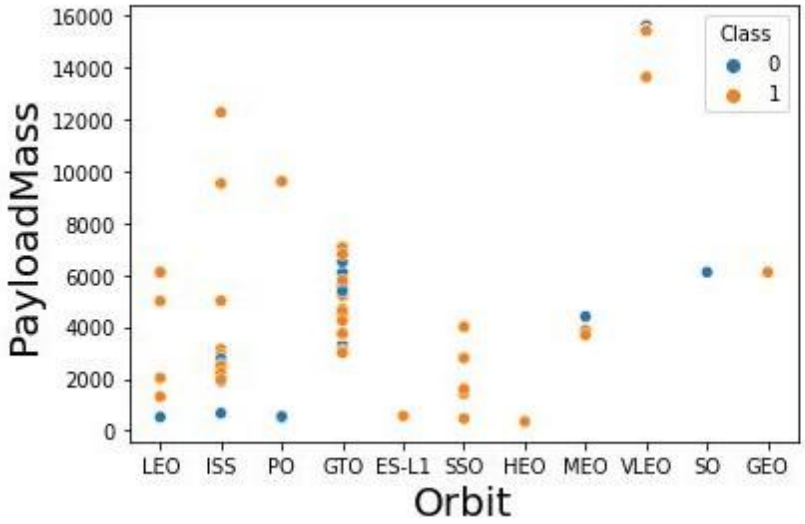
Flight No & Orbit



From the Visualization we can concluded that:

- Most Flight are to ISS,PO,GTO and VLEO
- MOST fails are for ISS,GTO
- SSO & VLEO has high success rate .

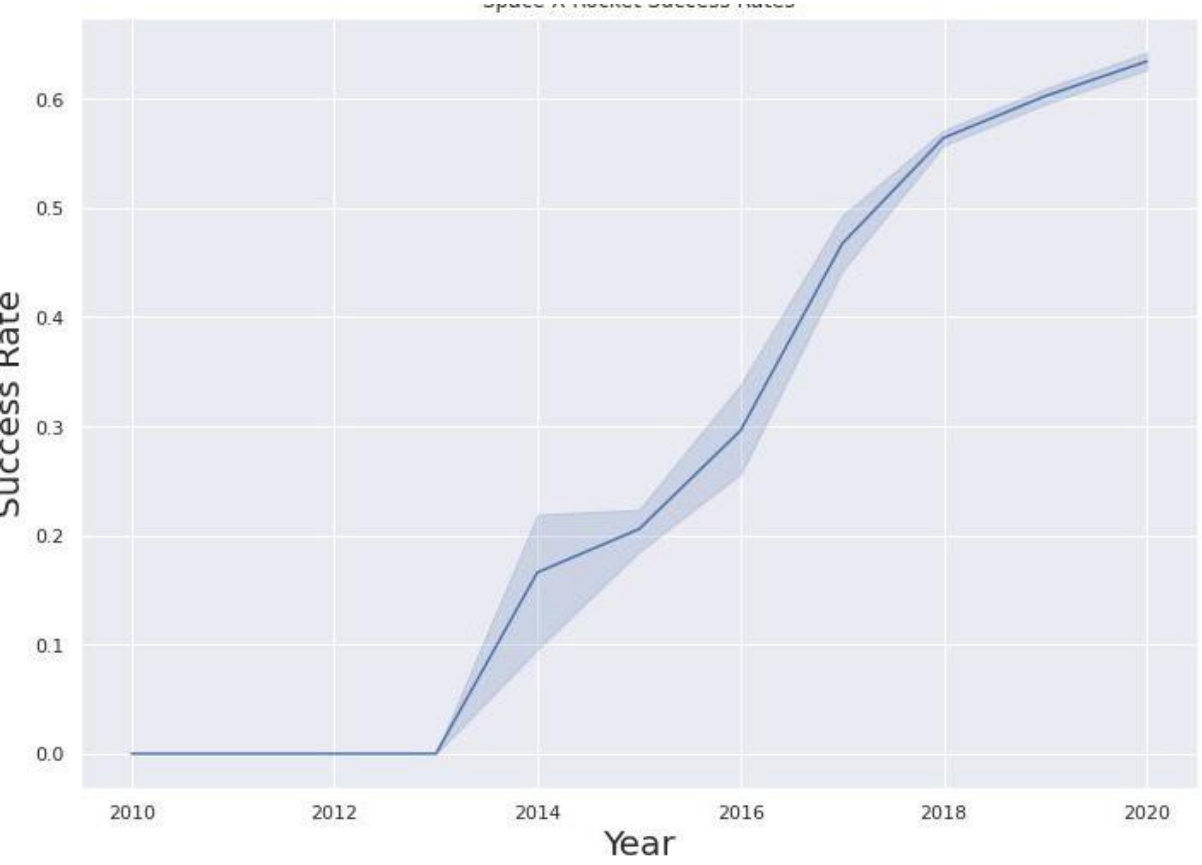
Payload & Orbit



From the Visualization we can concluded that:

- Higher Payload are to the VLEO
- Least Payload are for HEO,ISS,PO,ES-L1
- GTO has average payload size

Success Rate Trend



The rate of success of the launches increase over time since to the data collected from the previous fails and success launches .

Exploratory Data Analysis With Sql

Exploratory Data Analysis on the follow criteria:

Unique Sites

Max Payload

Average Payload

Day when First Success Landing

Success and Failures count

Boosters With Max Payload

EDA With Sql

For the categories above we find that :

Sites that SpaceX operates in are:

CCAFS LC-40,CCAFS SLC-40,KSC LC-39A,VAFB SLC-4E

Max Payload:48213

Average Payload for all Launches: 2928 Kgs

First Success Landing was Made on:06/05/2016

Booster Version that carry over 4000 kg and 6000 Kg :

F9 FT B1020,F9 FT B1022,F9 FT B1026,F9 FT B1021.2,F9 FT B1031.2

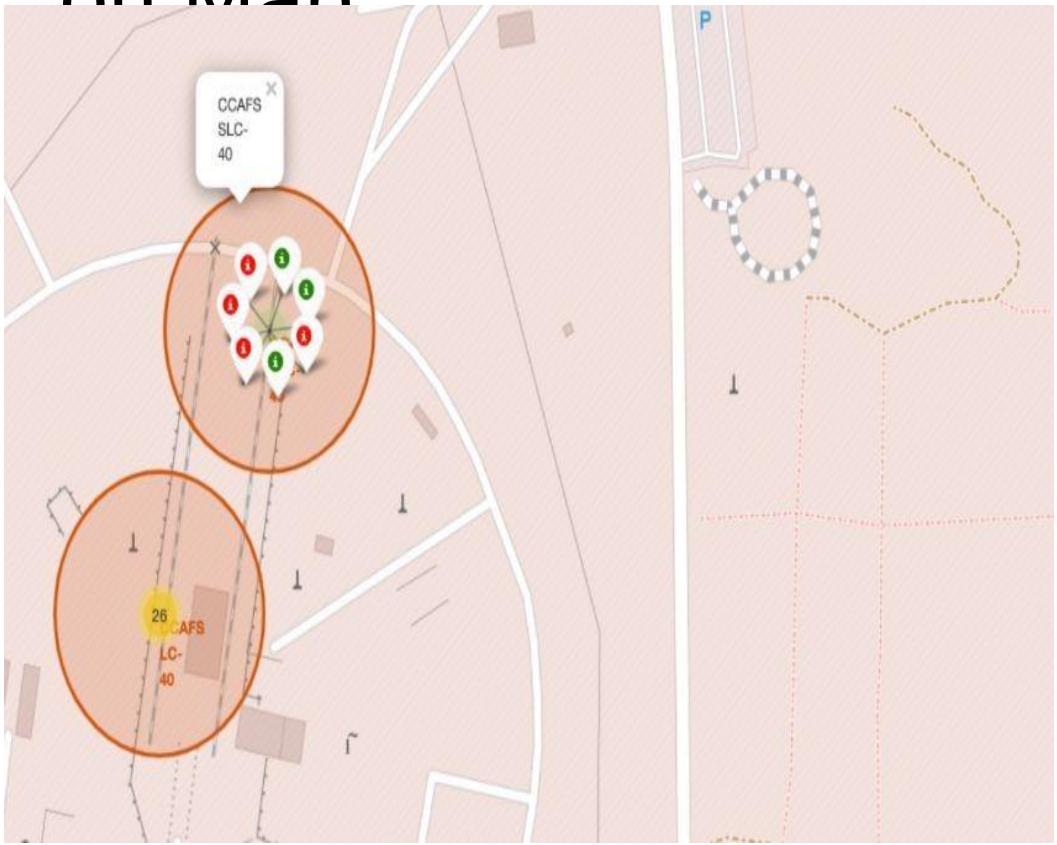
INTERACTIVE MAP WITH FOLIUM

Visualization of the launches for every site and every launch in a Interactive Map

Visualization of:

- Launch Sites
- Visualize the launches on the map base on Fail or Success

Visualize the Launches on Map



Data Set Contained 3 Separate Launch Sites that are displayed on the picture on the left .

This gives us insights to the launches success and failures .

PREDICTIVE ANALYSIS

Through Models,tuned for best performance we go the insights on the probability if a launch being success or a failure .

Models used include:

- **KNeighboursClassfier**
- **Decision Tree**
- **Logistic Regression**
- **Support Vector Machine**

Best Model Prediction

After Analyzing all the Models,the KNN was the best Model with accuracy of 77% and best Score of 87%

```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}

KNN = KNeighborsClassifier()
gscv=GridSearchCV(KNN,parameters,scoring="accuracy",cv=10)
KNN_cv=gscv.fit(X_train,y_train)
```

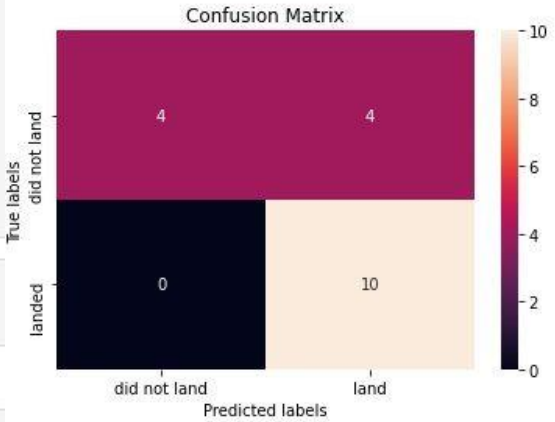
```
print("Accuracy",KNN_cv.score(X_test,y_test))
```

```
Accuracy 0.7777777777777778
```

```
print("tuned hpyerparameters :(best parameters) ",KNN_cv.best_params_)
print("accuracy :",KNN_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 4, 'p': 1}
accuracy : 0.8767857142857143
```

```
yhat = KNN_cv.predict(X_test)
plot_confusion_matrix(y_test,yhat)
```



True Positives

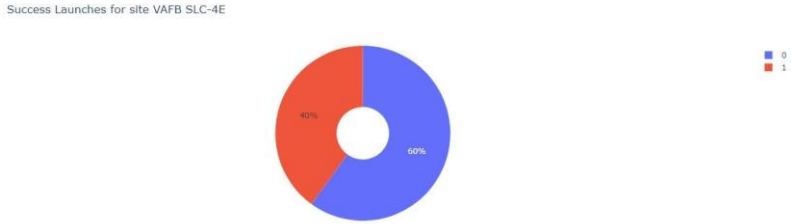
INTERACTIVE WITH DASH

Visualization of the Launches from Site in Dashboard

Visualization of:

- Success Launch Launch Sites
- Visualize payload from different sites with rangeSlider for interacting with the plot

Total Success Launches By all sites



Payload range (Kg):