Winning Space Race with Data Science

YouTian Peng
2021/12/30

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  After collecting the data, data pre-processing was implemented for further visual analysis, characterization, or for building machine learning classification models.
- Summary of all results

Decision Tree is the most accurate model for prediction among all. Base on final decision tree model which demonstrates a final chance of a successful landing of Falcon 9 about 73%

Introduction

- Project background and context

SpaceX 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 SpaceX can reuse the first stage.

- Problems you want to find answers

If the Falcon 9 first stage will land successfully. 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 SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Obtain Data through API
  - Obtain Data through WebScaping
- Perform data wrangling
  - Dealing with Missing Values
  - Convert Categorical Data for Featuring
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree, KNN

Data Collection

There are two different methods for data collection:

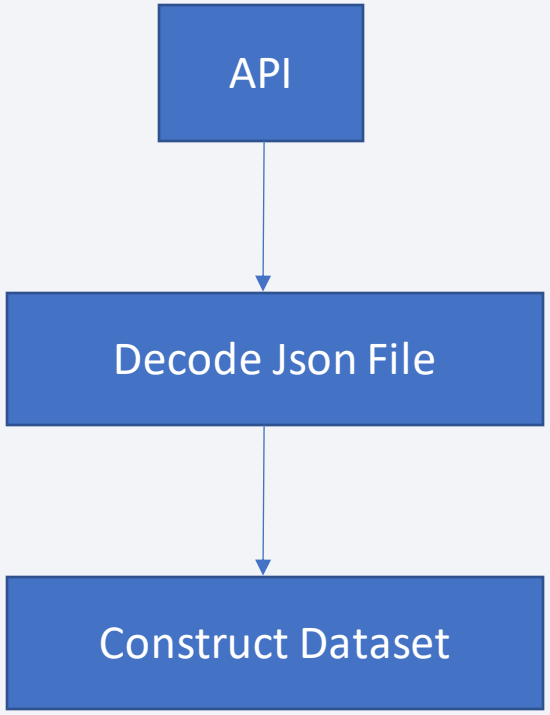
- Web API
- Web Scraping

Data Collection – SpaceX API

- Firstly I request and parse data using the GET request, then JSON file was decoded. After that, dataset was extracted for filtering "Falcon 9" launches data.

- GitHub URL:

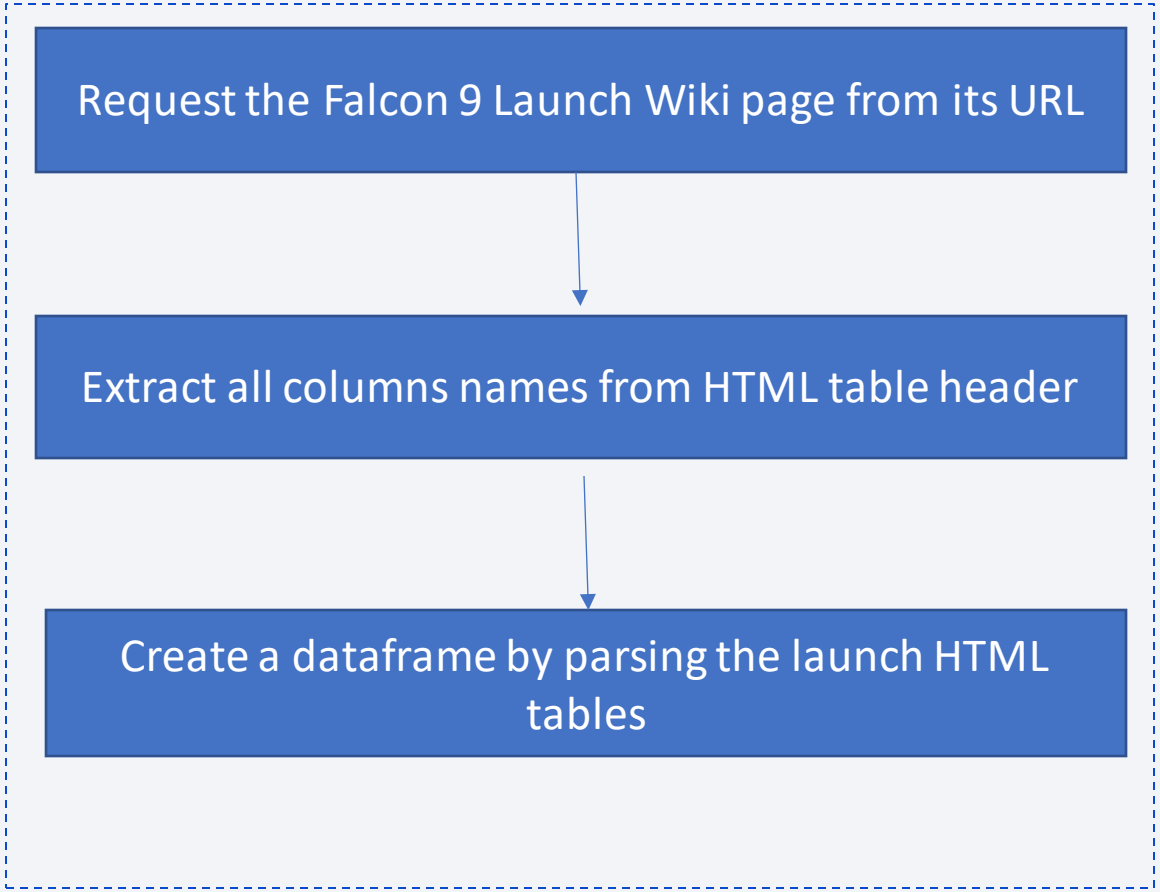
https://github.com/rorschachwilpeng/cousera_IBM_ds/blob/main/ds_capstone/Week1%20-%3E%20Introduction/Collecting%20the%20Data/Data%20Collection%20with%20API.ipynb



Data Collection - Scraping

- Firstly, I requested the Falcon 9 launch wiki page from its URL. Then all columns/variables names from HTML table header were extracted. After that, only useful headers were selected for creating dataframe.
- GitHub URL:

https://github.com/rorschachwilpeng/cousera_IBM_ds/blob/main/ds_capstone/Week1%20-%3E%20Introduction/Collecting%20the%20Data/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

Two main tasks of the Data Wrangling:

- Dealing with missing values

  Two column "PayloadMass", "LandingPad" exist missing values, only "PayloadMass" missing values were replaced with mean value of rest in this column. Data in the "LandingPad" retain NaN due to the lack of utilization of this data.

- Convert the "outcome"(landing outcome) into binary results.

  Data in outcome were converted into 1/0. "1" means landing successfully, where "0" means landing unsuccessfully.

GitHub URL:

https://github.com/rorschachwilpeng/cousera_IBM_ds/blob/main/ds_capstone/Week1%20-%3E%20Introduction/Data%20Wrangling/Data%20Collection%20with%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Three types of chars were drew: "Scatter Point Chart", "Bar Chart", "Line Chart"

1. Scatter Point Chart: For observing each pair of attributes relative relationship with landing result.
2. Bar Chart: For observing the relationship between landing success rate and orbit types.
3. Line Chart: For observing the relationship between landing success rate and years.

- GitHub
  URL: https://github.com/rorschachwilpeng/cousera_IBM_ds/blob/main/ds_capstone/Week2%20-%3E%20EDA%20with%20SQL/Exploratory%20Analysis%20Using%20Pandas%20and%20Matplotlib/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

GitHub URL: https://github.com/rorschachwilpeng/cousera_IBM_ds/blob/main/ds_capstone/Week2%20-%3E%20EDA%20with%20SQL/Exploratory%20Analysis%20Using%20SQL/EDA%20with%20SQL.ipynb

- Display the names of the unique launch sites in the space mission

```
%sql select distinct(launch_site) from SPACEXTBL;
```

- Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;
```

- Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';
```

- Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version like 'F9 v1.1%';
```

- List the date when the first successful landing outcome in ground pad was acheived.

```
%sql select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)';
```

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql select distinct(booster_version) from SPACEXTBL where landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000;
```

- List the total number of successful and failure mission outcomes

```
%sql select mission_outcome,count(*) as Counter from SPACEXTBL group by mission_outcome;
```

- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select distinct(booster_version) from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);
```

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select landing__outcome,booster_version,launch_site,date from SPACEXTBL where Date like '2015%' and landing__outcome = 'Failure (drone ship)';
```

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql select landing__outcome,count(landing__outcome) as count from SPACEXTBL where DATE BETWEEN '2010-06-04' and '2017-03-20' GROUP BY landing__outcome ORDER BY count(landing__outcome) DESC;
```

Build an Interactive Map with Folium

There are three objected be added to Map:

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

- GitHub URL:

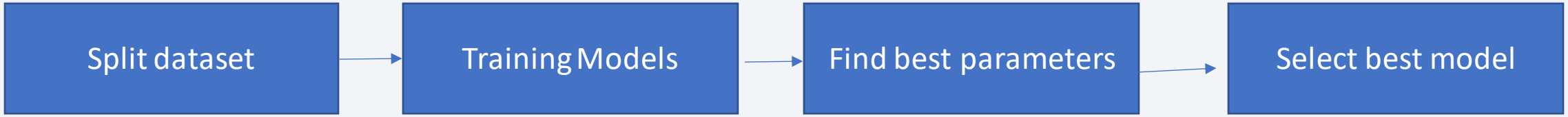
https://github.com/rorschachwilpeng/cousera_IBM_ds/blob/main/ds_capstone/Week3%20-%3E%20Interactive%20Visual%20Analytics%20and%20Dashboard/lab_jupyter_launch_site_location.ipynb

Predictive Analysis (Classification)

- Four main steps for performing best model:

1. Split train and test datasets;
2. Trained Models(Logistic Regression, SVM, Decision Tree, KNN)
3. Find best parameters for each model with GridSearch CV
4. Select best model base on prediction accuracy

Flowcharts:



- GitHub URL:

https://github.com/rorschachwilpeng/cousera_IBM_ds/blob/main/ds_capstone/Week4%20-%3E%20Predictive%20Analysis%20(Classfication)/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

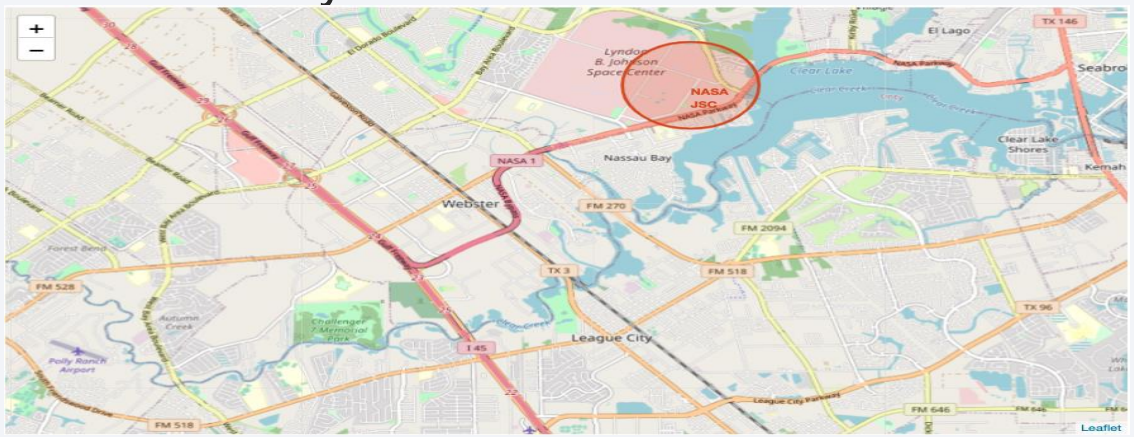
Results

- Exploratory data analysis results

1. At the VAFB-SLC launch site there are no rockets launched for heavy payload mass(greater than 10000).
2. "ES-L1", "GEO", "HEO", "SSO" orbit type have highest success rate(100%)
3. With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS

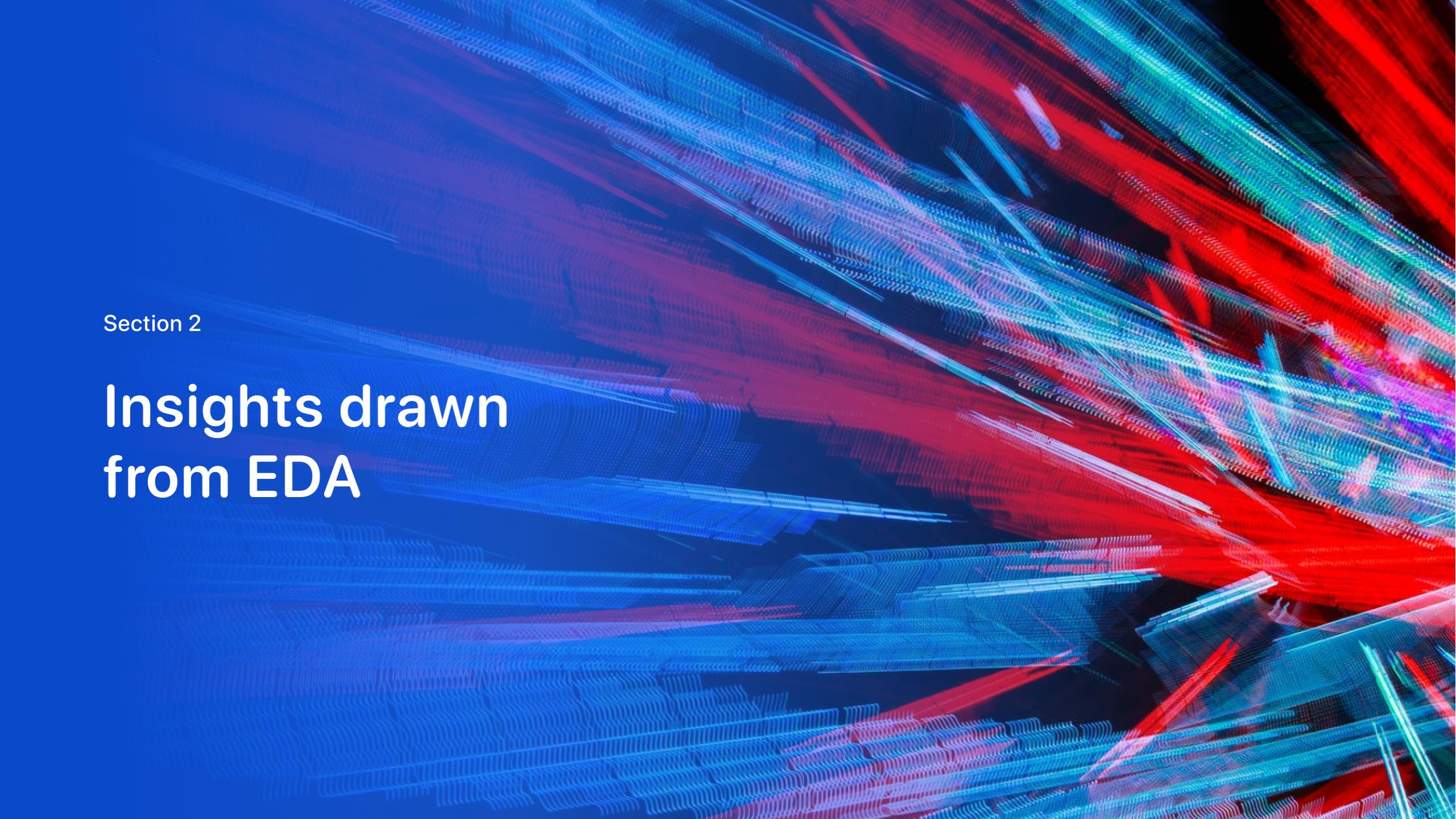
…

- Interactive analytics demo in screenshots



- Predictive analysis results

1. Desicision Tree is best model;
2. The final prediction for landing success rate is about 73%

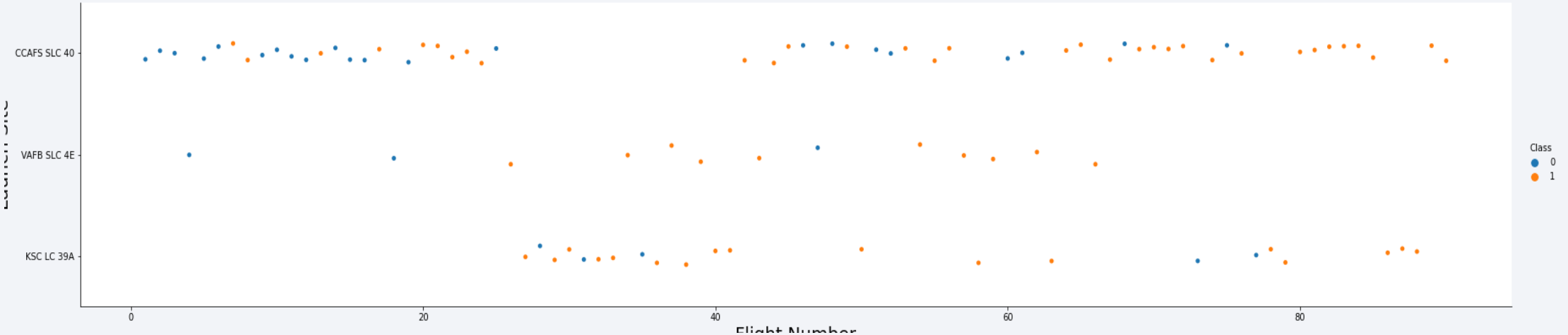


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- Scatter plot of Flight Number vs. Launch Site



- Explanation for the plot:

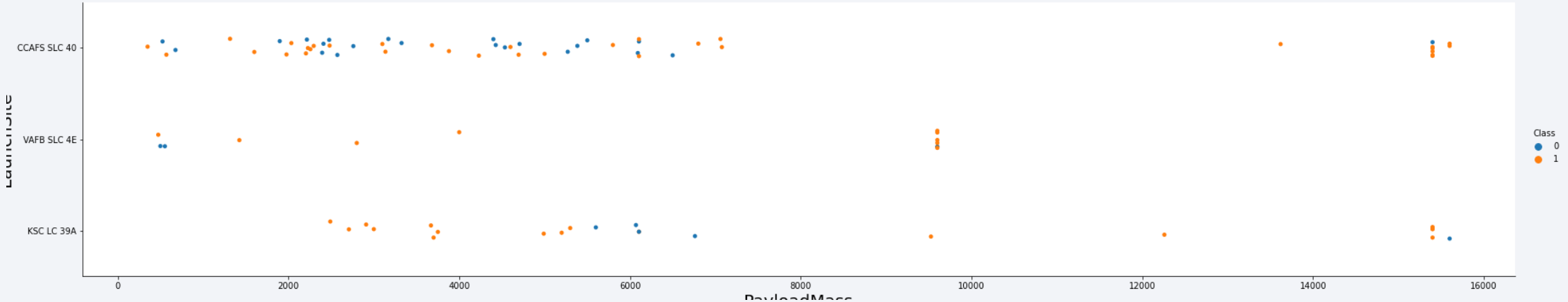
  For "CCAFS SLC 40": The success rate rise with the increasement of the flight numbers;

  For "VAFB SLC 4E": The success rate rise with the increasement of the flight numbers;

  For "KSC LC 39A": There is no instance when flight number is smaller than 28.

Payload vs. Launch Site

- Show a scatter plot of Payload vs. Launch Site

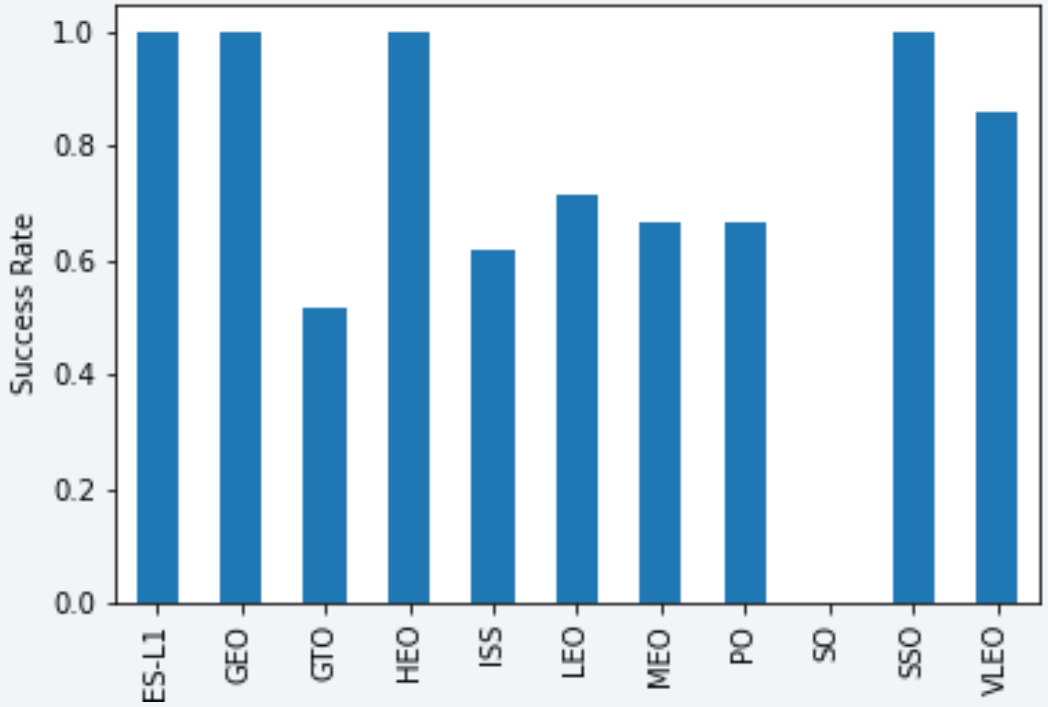


- Plot Explanations:

The VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

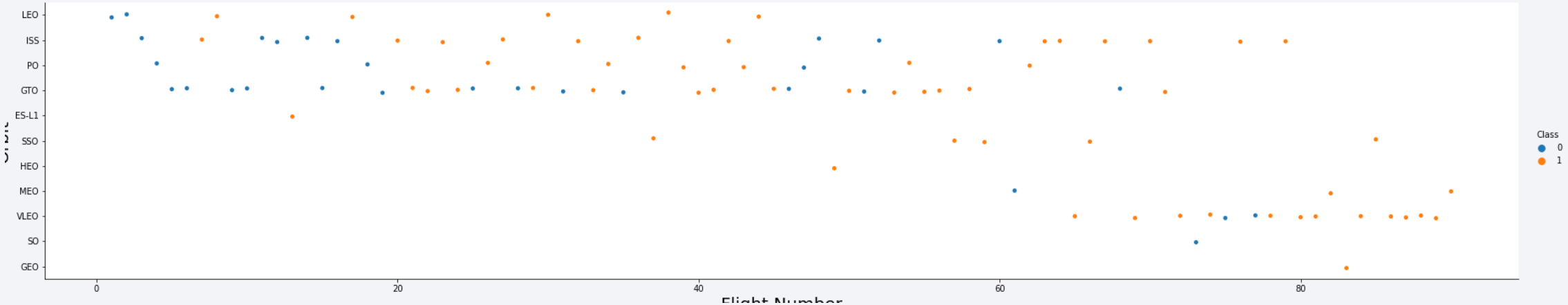
- Show a bar chart for the success rate of each orbit type



- Plot explanations: Orbit type of "ES-L1", "GEO", "HEO", "SSO" have highest success rate.

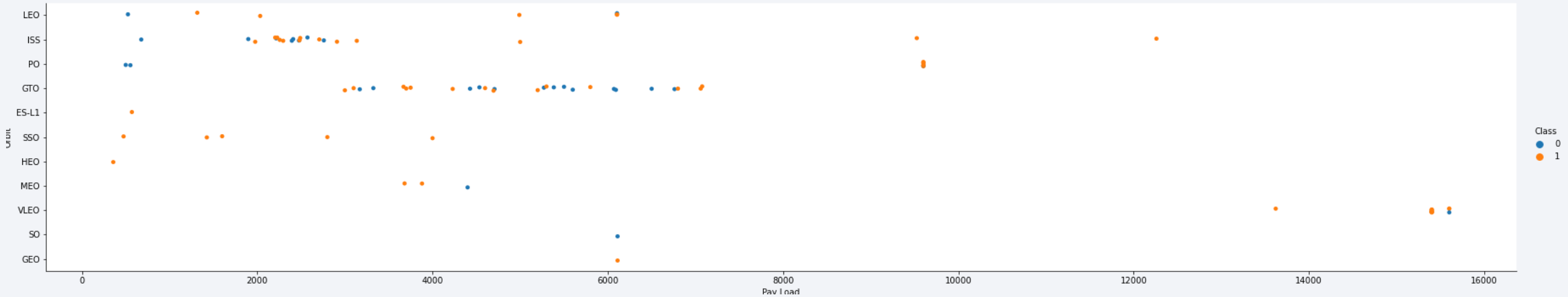
Flight Number vs. Orbit Type

- Show a scatter point of Flight number vs. Orbit type



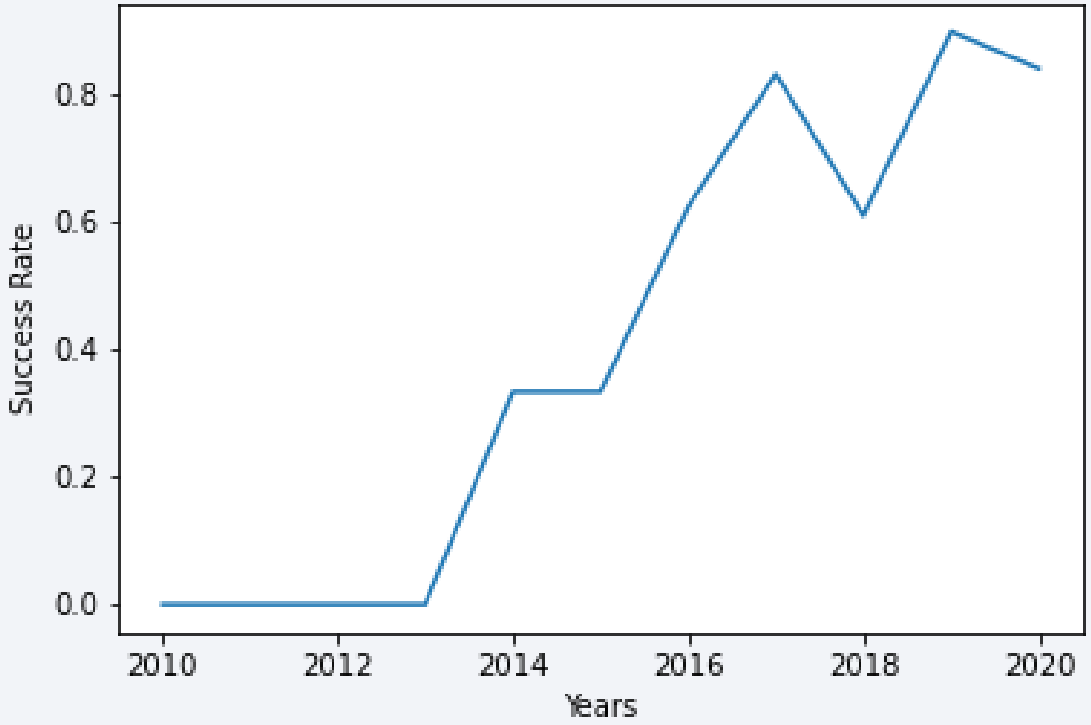
- Plot Explanations: You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- Plot Explaination: With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- Plot Explaination: You can observe that the sucess rate since 2013 kept increasing till 2020

All Launch Site Names

- Find the names of the unique launch sites
- Present your query result with a short explanation here

Query Result: %sql select distinct(launch_site) from SPACEXTBL;

Explanation: ...

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
- Present your query result with a short explanation here

Query Result: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5;

Explanation: ...

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
- Present your query result with a short explanation here

Query Result: %sql select sum(payload_mass__kg_) from SPACEXTBL where customer = 'NASA (CRS)';

Explanation: Calulate the total payload with "sum()" function then select boosters from NASA base on condition "where customer = 'NASA(CRS)' ".

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
- Present your query result with a short explanation here

Query Result: %sql select avg(payload_mass__kg_) from SPACEXTBL where booster_version like 'F9 v1.1%';

Explanation: Reach bosster version with fuzzy searches.

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
- Present your query result with a short explanation here

Query Result: %sql select min(DATE) from SPACEXTBL where landing__outcome = 'Success (ground pad)';

Explanation: ...

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

Query Result: %sql select distinct(booster_version) from SPACEXTBL where landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ > 4000 AND payload_mass__kg_ < 6000;

Explanation: ...

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Query Result: %sql select mission_outcome,count(*) as Counter from SPACEXTBL group by mission_outcome;

Explanation: ...

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Present your query result with a short explanation here

Query Result: %sql select distinct(booster_version) from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL);

Explanation: ...

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

- Present your query result with a short explanation here

Query Result: %sql select landing__outcome,booster_version,launch_site,date from SPACEXTBL where Date like '2015%' and landing__outcome = 'Failure (drone ship)';

Explanation: ...

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- Present your query result with a short explanation here

Query Result: %sql select landing__outcome,count(landing__outcome) as count from SPACEXTBL where DATE BETWEEN '2010-06-04' and '2017-03-20' GROUP BY landing__outcome ORDER BY count(landing__outcome) DESC;

%sql SELECT * FROM SPACEXTBL where DAYNAME(DATE)='Friday' LIMIT 5 ;

Explanation: ...

Section 4

Launch Sites Proximities Analysis

Mark all launch site on a map

- All four launch sites including "CCAFS LC-40", "CCAFS SLC-40", "KSC LC-39A", "VAFB SLC-4E" were marked on the map base on their coordinates.

Mark the success/failed launches for each site on the map

- "KSC LC 39-A" have highest success rate
- "CCAFS SLC-40" have highest failed rate

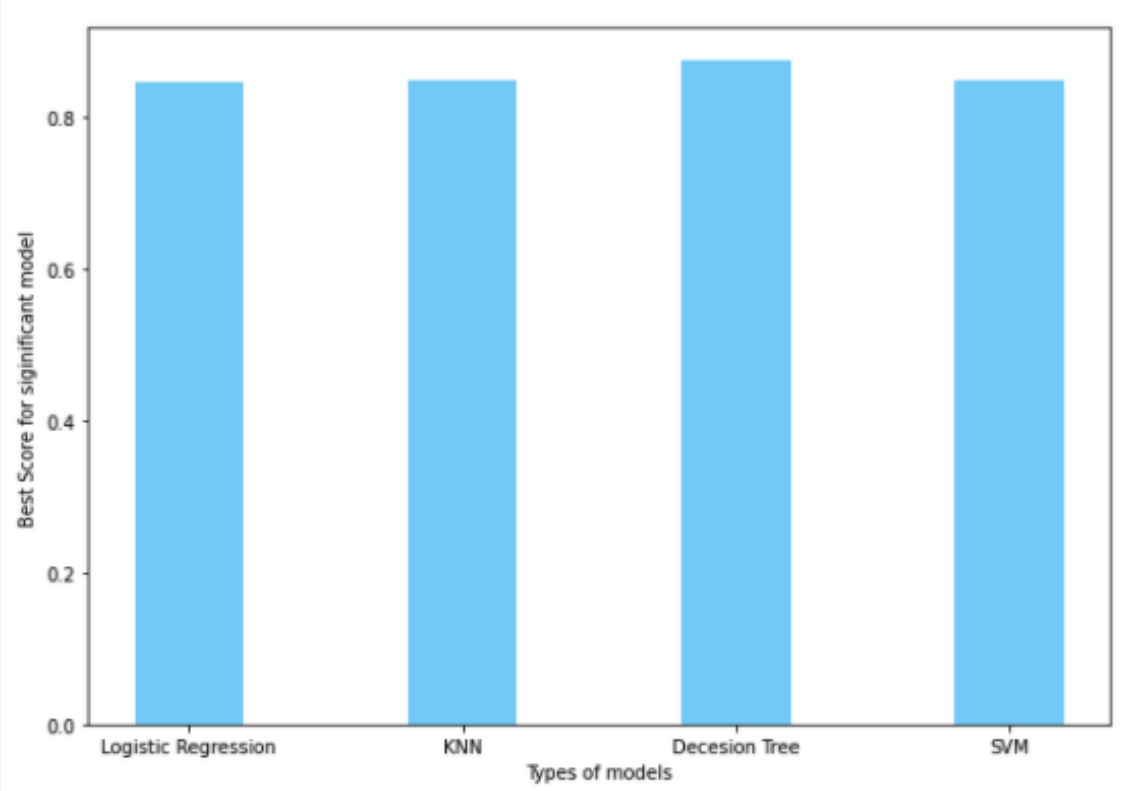
TASK 3: Calculate the distances between a launch site to its proximities

- "CCAFS LC-40", "CCAFS SLC-40" are close to the "Centaur Road" and costline;
- "KSC LC-39A" is close to the "Merrat Island National Wildlife Refuge";
- "VAFB SLC-4E" is close to the "Dix Road".

Section 6

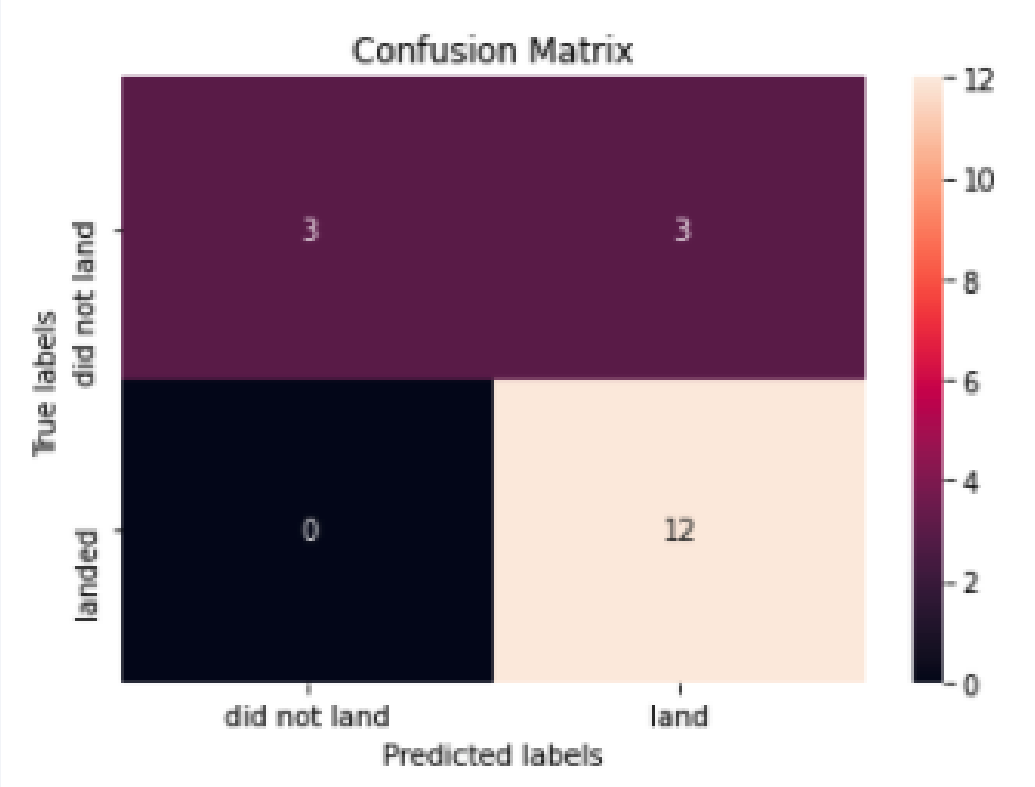
Predictive Analysis (Classification)

Classification Accuracy



- Decision Tree have highest clasification accuracy.

Confusion Matrix



- Base on the confusion matrix, all "landed" label are predicted correctly. Whears, for "did not land", half of label are predicted correctly.

Conclusions

- EDA: Among all orbit type, "ES-L1", "GEO", "HEO", "SSO" have high success rate
- EDA: The success rate since 2013 kept increasing till 2020
- EDA: The attributes of processed datasets including: 'FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins', 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial';
- Classfication: Decision Tree is best model with highest classfication accuracy.

Thank you!