SPACE X FALCON 9
FIRST STAGE LANDING PREDICTION

Capstone Project for 'IBM Professional Certificate in Data Science with Python

- Huseyn Panahov

INTRODUCTION

- In this project, I constructed a machine learning model that can predict whether Falcon 9 will land successfully in the first stage. Falcon 9 is classified as a medium-lift partially reusable rocket, used to launch hefty communications and satellites into Earth orbit or ferry austronaouts to and from the International Space Station.
- As of April 2022, SpaceX offers Falcon 9 rocket launches for USD 62 million, which means around USD 1,200 per pound of payload. For comparison, per pound cost of SpaceX competitors is 3 to 5 times more expensive, whereas traditional NASA space shuttles, retired in 2011, cost an average of USD 1.6 billion per flight. SpaceX is able to provide rocket launches for unprecedented low prices, because it can reuse the first stage, which significantly reduces the demand for new cores.
- Therefore, determining whether the first stage will land, helps to estimate the cost of a launch. This information can also be used if another company wants to bid against SpaceX for a rocket launch.

EXECUTIVE SUMMARY

- I use API requests to pull data from < https://api.spacexdata.com/v4 > and web-scraping to collect data from Wikipedia.
- I use Machine Learning methods to construct a predictive model, which achieves 83% accuracy on the test data.
- It means our model can accurately predict the outcome of15 out of every 18 flights

DATA COLLECTION

- Primary data collection method was API request:

  https://api.spacexdata.com/v4

- Additionally, web-scraping methodology was to collect data from Wikipedia:

  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

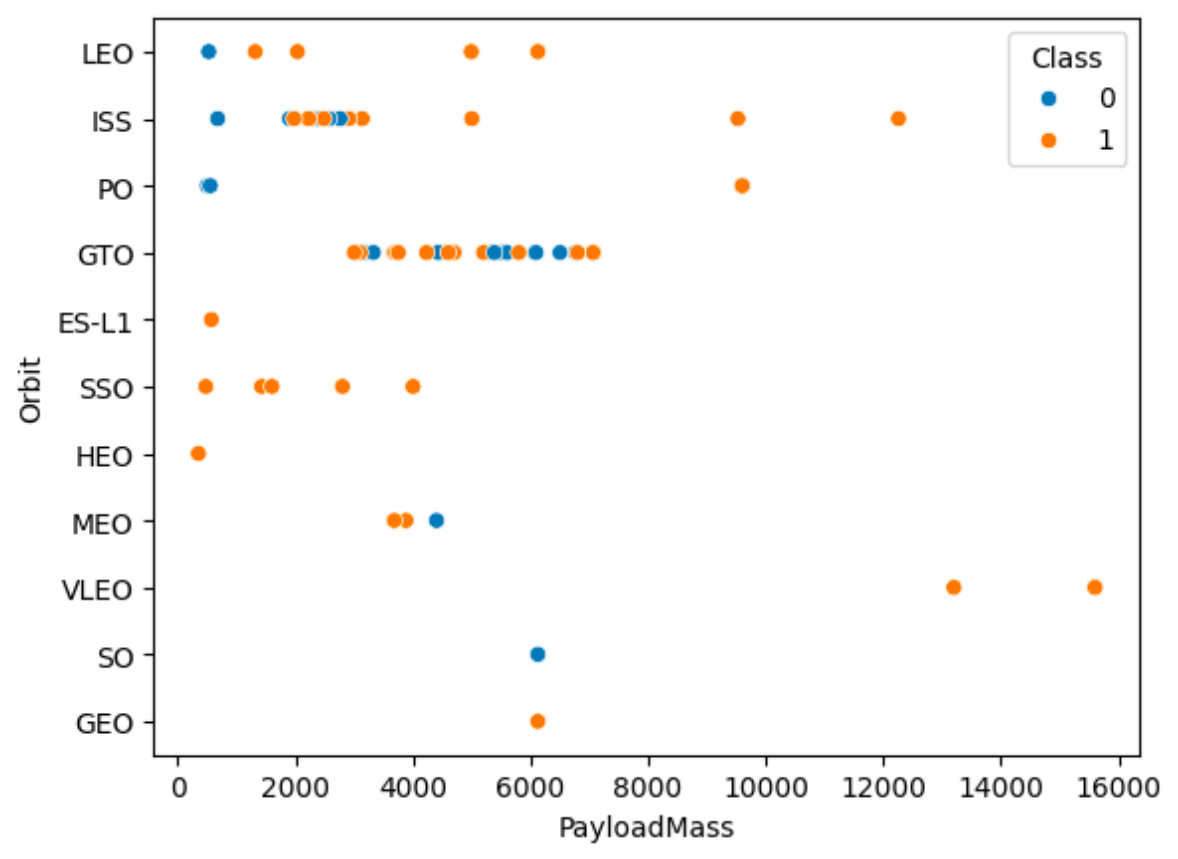
DATA WRANGLING

- Data cleaning
- Created the outcome variable

  - new Class variable

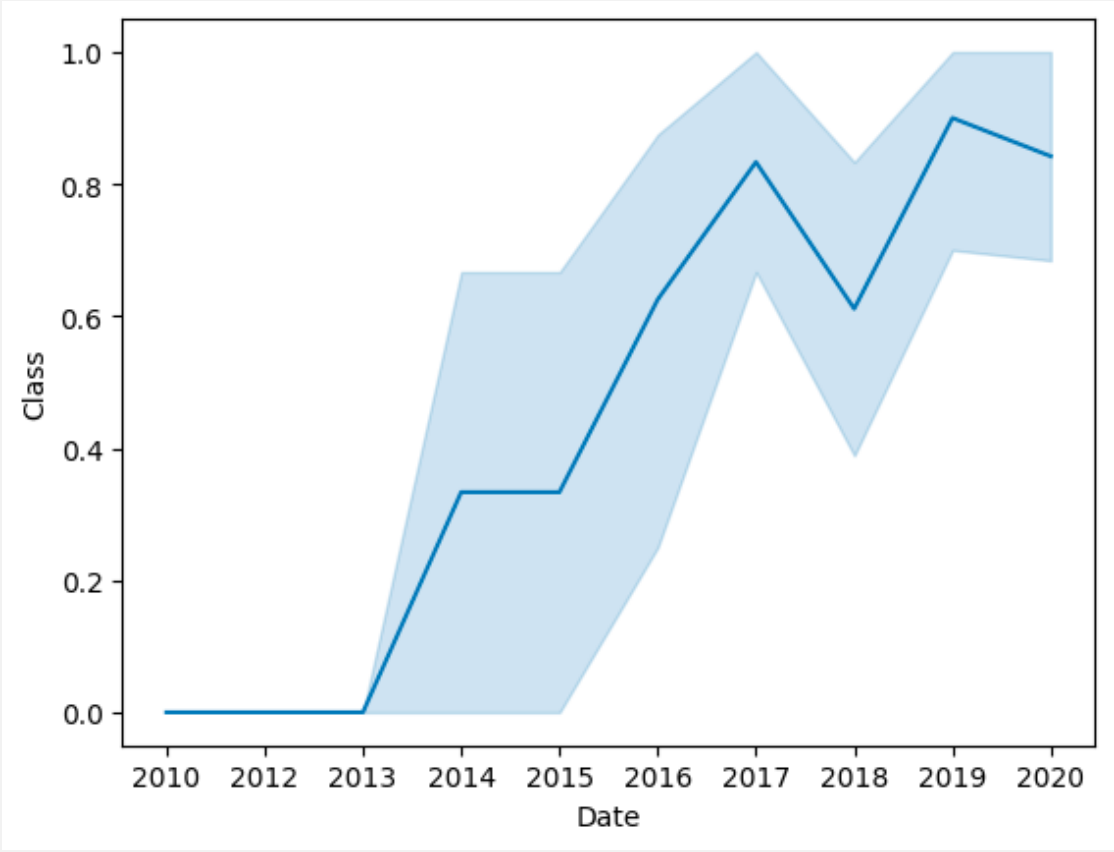
Landing_outcomes	Class
None None	0
Fals ASDS	
False Ocean	
None ASDS	
False RTLS	
True ASDS	1
True RTLS	
True Ocean	

DATA EXPLORATION



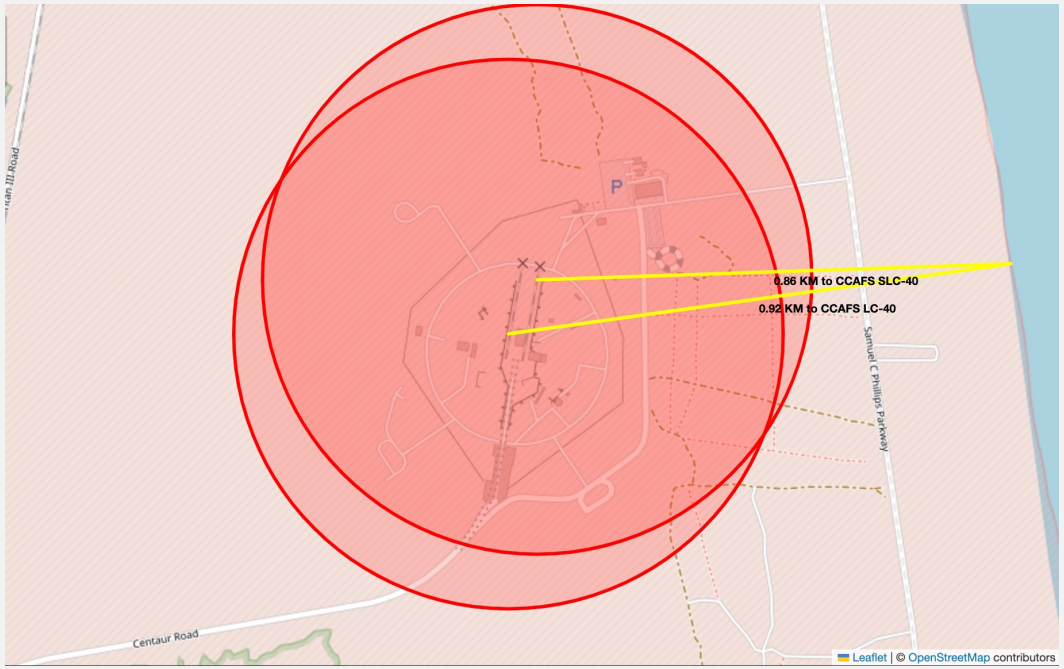
We notice that flights with payload mass higher than 8000 are more likely to be succesful than flights with payloads under 2000

DATA EXPLORATION

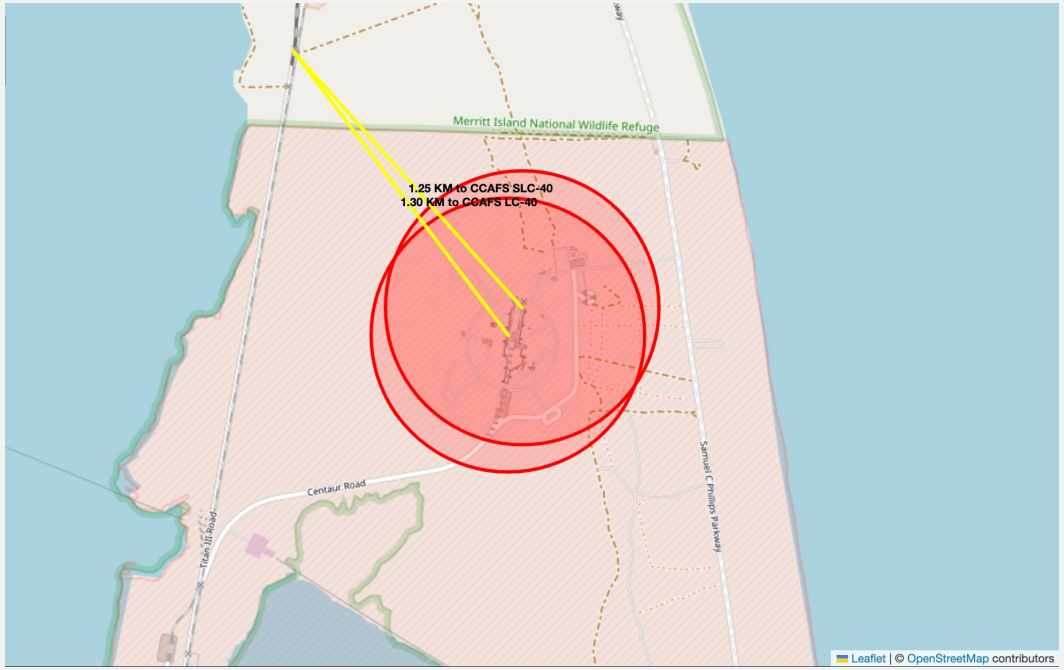


Overall success rate of the flights has bee consistently increasing since 2013, absent a slight dive in 2018.

INTERACTIVE MAPS USING FOLIUM



Proximity to coastline



Proximity to raiway

METHODOLOGY

- By now, we have preliminary insights about which variables have the most significant impact on the landing outcomes
  - FlightNumber', 'PayloadMass', 'Orbit', 'LaunchSite', 'Flights', 'GridFins',
  - 'Reused', 'Legs', 'LandingPad', 'Block', 'ReusedCount', 'Serial'
- We will select these features and use them in constructing the predictive model
- Use ‘get_dummies’ function and create a new features dataframe to apply OneHotEncoder to these variables.
- Save the new dataframe with 83 variables as a separate csv file

METHODOLOGY

- Split the data into training and testing data using the function train_test_split.
- The training data is divided into validation data, a second set used for training data; then the models are trained and hyperparameters are selected using the function GridSearchCV

CONSTRUCT ML MODELS

Training Model	Training Data Accuracy	Test Data Accuracy
Logistic Regression	0.846	0.833
Support Vector Machine	0.836	0.833
Decision Tree Clasifier	0.903	0.778
K-Nearest Neighbor Classifier	0.848	0.833

CONCLUSION

- We notice all our models, perform similarly on the test data, absent the decision tree model.
- With all things considered, it is best to choose the Logistic Regression model, because of its simplicity.
- It achieves 83% accuracy on the test data. It means our model can accurately predict the outcome of15 out of every 18 flights
- Overall, we had 90 flight registrations in our data set. With more flight recordings down the road we can further increase/improve the accuracy of our model.