CAPSTONE DATA SCIENCE PROJECT

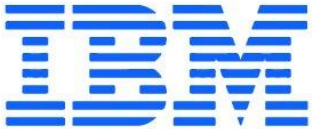
Michael Komer 06/2023

EXECUTIVE SUMMARY

This Summary Report is for IBM Data Science Professional – Certificate – Data Science Capstone Course

Some of the things covered included:

- Data Wrangling, best practices and tools
- Exploratory Data Analysis SQL
- Data Visualization, various tools and formats
- Python Data Analysis
- Machine Learning Models and Visualizations using various Python libraries: Pandas, Seaborn, Scikit Learn, Matplotlib
- Predictive Analysis methodologies and tools available



INTRODUCTION

My Name is Mike Komer, I will briefly share my background and the purpose of this report

- I currently work as an analyst and tableau data developer on the Enterprise Applications and Data Governance Team

- I oversee the Business Intelligence Applications like Tableau

- My goal was to complete this IBM Data Science Certification and further develop my toolset and methodologies, as well as show create a way to showcase them

- In this Capstone Course I will be sharing various course materials from prior sessions I was apart of to both showcase the newly added skills and tools from this certificate and on this report

DATA WRANGLING

Data Collection and Data Wrangling Process,
Key Take Aways:

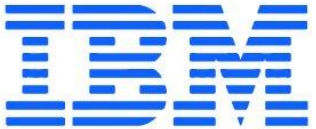
See the attached “Data Wrangling” Python lab coursework for full details and data work:

Data Wrangling Project Github

Performed Exploratory Analysis on SpaceX Launch Data to find data patterns, anomalies and determine the label for training supervised models

Imported Various Libraries Python Libraries to build visualizations during the exploratory phase; seaborn, Scikit Learn, matplotlib, pandas

Custom Calculated Fields for Launch Data, Training Label Definition, and data wrangling best practices



INTERACTIVE VISUAL ANALYTICS

Interactive Visualizations:

Interactive Visual Analytics

Some of my favorite visuals made during certificate program displayed on the right.

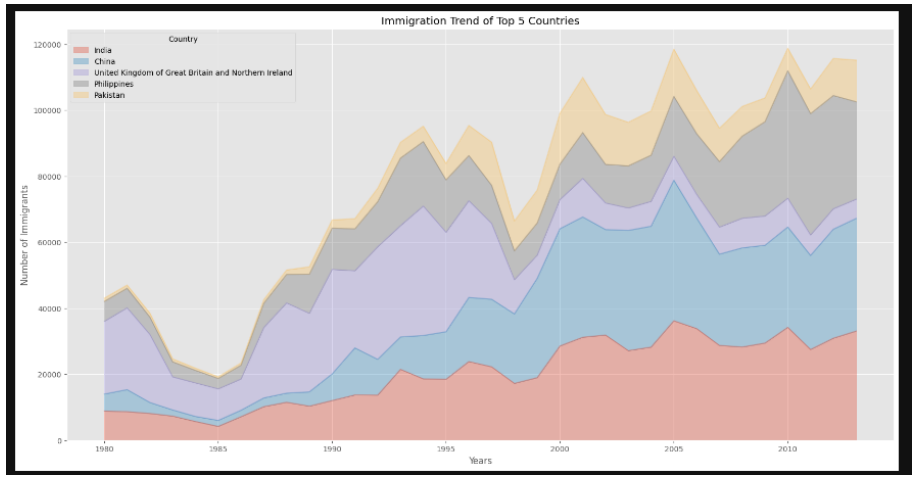
Performed Exploratory Data Analysis on SpaceX Launch Data to find data patterns, anomalies and determine the label for training supervised models

Explored different visualization types and use cases with SpaceX dataset

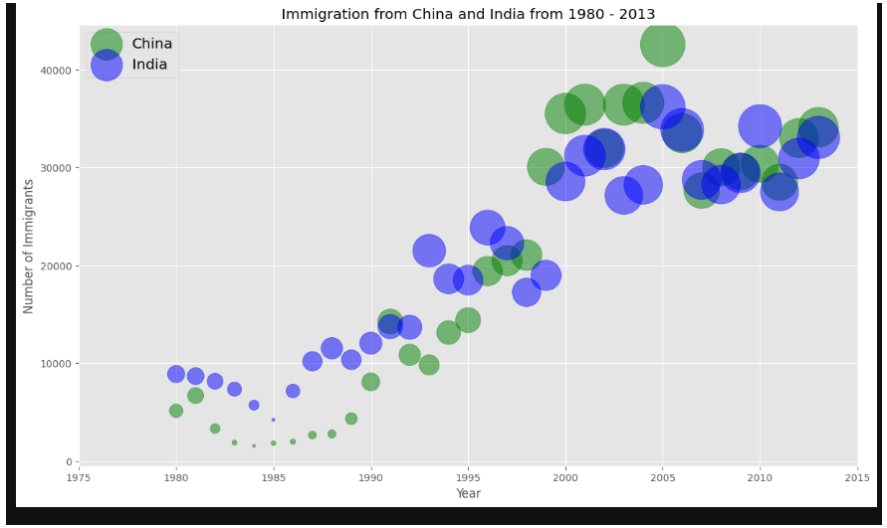
Began Classification Modeling and Testing strategy

Imported Various Libraries Python Libraries: seaborn, Scikit Learn, matplotlib, pandas

Stacked Area Plot Made During Certificate Program



Bubble Plot Time Series Data



PREDICTIVE ANALYSIS

EDA And Predictive Viz Project
Key Take Aways:

See the attached "Exploratory Data Analysis.." github project work

Exploratory Data Analysis and Advanced Data Visualization

Real Case Study Analysis on SpaceX data and beginning look at how different data methods can be applied, below:

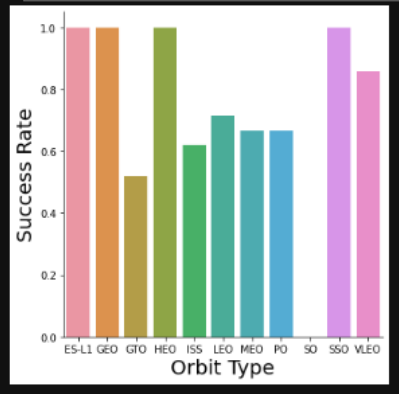
LEFT: Orbit Type Relationship Success

TASK 3: Visualize the relationship between success rate of each orbit type

Next, we want to visually check if there are any relationship between success rate and orbit type.

Let's create a `bar chart` for the sucess rate of each orbit

In [6]:
```
# HINT use groupby method on Orbit column and get the mean of Class column
sns.catplot(x= 'Orbit', y = 'Class', data = df.groupby('Orbit')['Class'].mean().reset_index(), kind = 'bar')
plt.xlabel('Orbit Type',fontsize=20)
plt.ylabel('Success Rate',fontsize=20)
plt.show()
```

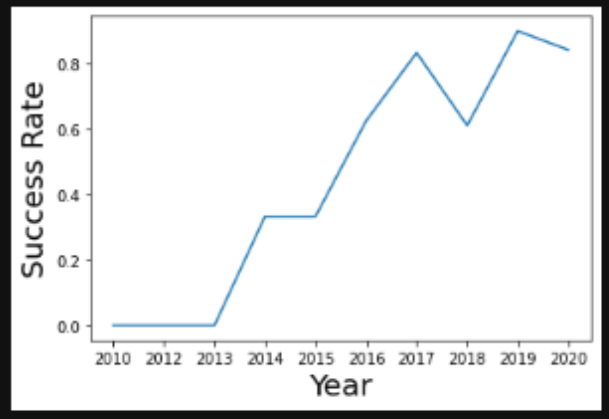


Analyze the ploted bar chart try to find which orbits have high sucess rate.

RIGHT: Launch Success Trends

In [10]:
```
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
years = df.groupby(Extract_year(df['Date'])).mean()['Class']

sns.lineplot(x = years.index, y = years)
plt.xlabel('Year', fontsize = 20)
plt.ylabel('Success Rate', fontsize = 20)
plt.show()
```



you can observe that the sucess rate since 2013 kept increasing till 2020

EDA WITH VISUALIZATION

Exploratory Data Analysis and Visulization
Key Take Aways:

See the attached "Predictive Analysis and Machine Learning Lab github project work

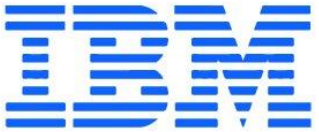
Predictive Analysis and Machine Learning Lab

Real Case Study Analysis on SpaceX data and Machine Learning Prediction Models

Standardized the dataset, split data into groups, and began various data modeling tests and visual analysis on dataset

Tested various Machine Learning Models against dataset and applied different visualization tools and analysis strategies

Determined the machine learning method best fit for current data across Logistic Regression, Classification Trees, and Support Vector Management Models



EXPLORATORY DATA ANALYSIS

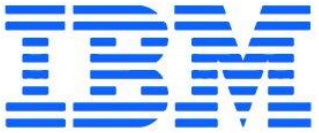
EDA And Advanced Data Viz Project
Key Take Aways:

See the attached "Exploratory Data Analysis.." github project work
Exploratory Data Analysis and Advanced Data Visualization

Real Case Study Analysis on SpaceX data and beginning to look at an indepth data exploration process, what that entails, and feature engineering

Came up with Visual Analysis Methodologies;
(Case Study Example: Task 2 Visualize relationship between SpaceX Payload and Launch Site)
- this was a to see how it might be applied in a real-world situation that made the project more interesting and encouraging to patriciate in

Connected to SQL Database Server, reviewed various SQL tricks



FOLIUM INTERACTIVE MAP

Interactive Visualizations

Interactive Visual Analytics

Focused on the Map Data Launch Analysis, added elements to the maps within Python that displayed in depth details of launch sites, activity levels, trends in graph visualization marks

PLOTLY VIZ DASHBOARD RESULTS

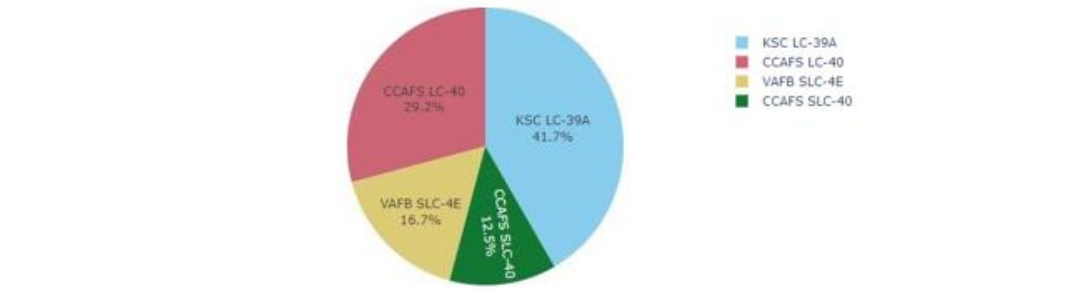
Plotly Dashboard:

Plotly Viz Link

Continuation of SpaceX Case Study,

Dashboard is made up of pie chart and scatter plots

Pie chart displays distribution of successful landing and to visualize launch site success rates where the Scatter plot helps show the varying factors to success across launch sites



PREDICTIVE ANALYSIS CLASSIFICATION

Predictive Analysis(Classification):

Predictive Analysis Classification

Continuation of SpaceX Case Study

Best Performing Methodology:

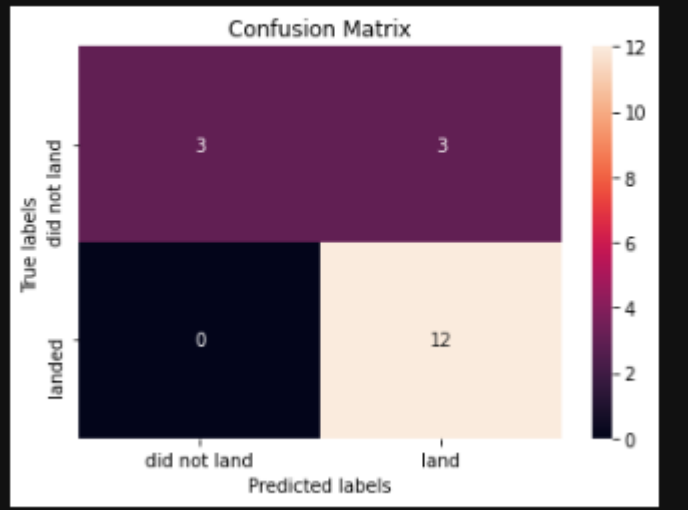
Out[31]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Tree Model considered most popular and shown here yields best accuracy

Confusion Matrix Summary

In [29]:
```
knn_yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test, knn_yhat)
```



INNOVATIVE INSIGHTS:

- Since all models performed the same for the test set, the confusion matrix is the same across all models.
- The models depict 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).
- The models appear to over predict successful landings, possibly in part due to sample size.

CONCLUSION

FINAL THOUGHTS:

Overall satisfied with IBM Data Science Professional Certificate Program

Was able to test out new tools and experiment with different visualization models and strategies in a semi structured self paced environment

Expanded on my Jupyter Notebook, GitLab, Python, Python Libraries, Python Visualization, and Machine Learning fundaments

Was also nice to have real case studies and practical use scenarios attached to the lessons, made it easier to stay engaged.