Prediction for Best Airlines

Group 01

Tomasz Olewicz - Data Engineer
Alex Lamp - Data Visualization Specialist
Irene Kang - Machine Learning Specialist
Joe Chun - Just Joe Chun

TABLE OF CONTENTS

“The World is one big data problem.”

— Andrew McAfee, co-director of the MIT Initiative

01
Our Topic

Our project's goal :

- Classification of each airline that traffics through San Francisco International Airport(SFO)
- Train & test each airline with different machine learning models to achieve great accuracy and finally use them to predict their performance
- Provide precise information of each airline's quality to potential future customers, therefore they can have smarter decision for their own benefit

Which airline would you choose ?

We chose this because

Popularity

A lot of people in the world travel for business, vacation, reunion, etc.

San Francisco International Airport(SFO) is the point of intersection and the major gateway to the rest of the world

We plan to achieve our goal by

Machine Learning Model

Build a machine learning model to predict which airline has better quality so more potential customers in the future can have better experience

02

Source of Dataset

Which Dataset did we use to achieve our goal?

SFO Air Traffic Passenger and Landings Statistics

https://www.kaggle.com/san-francisco/sf-air-traffic-passenger-and-landings-statistics?select=air-traffic-landings-statistics.csv

Top 100 Airline Fleets

https://www.kaggle.com/traceyvanp/airlinefleet

Airline Performance

airline_performance.csv

includes information of each airline's traffic details such as

- Passenger count
- Landing Count
- Number of Airplanes
- Total Cost
- Average Fleet Age
- Each Published Airlines

03

Tools for the Project

Which tools did we use on the project?

Pandas for

- Encoding
- Trimming out unnecessary values
- Merging different frames together
- Graphs, plots, 3D modeling
- Outcome & Further analysis plan

Google Slides & Tableau for

- Visualization
- Storytelling
- Presentation

Machine Learning Model

- Logistic Regression
- Support Vector Machine(SVM)
- Scikit Library

04

Data Exploration and Analysis

Three different datasets were analyzed and prepared for execution in the ML model

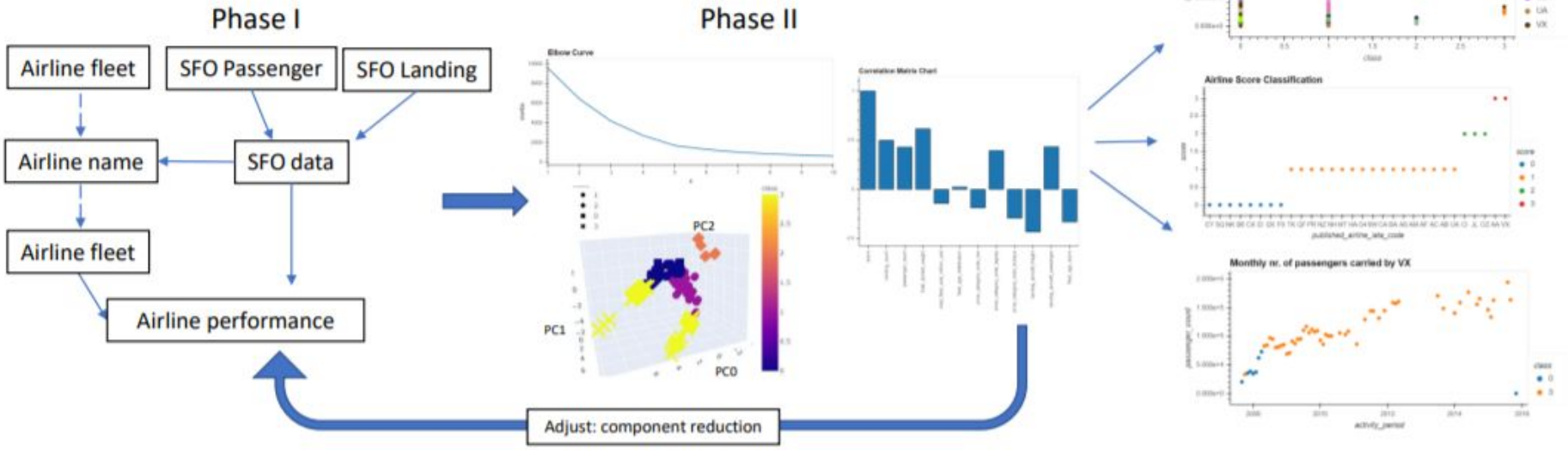
Using Python-pandas : cleaned data frames and further analysis was done

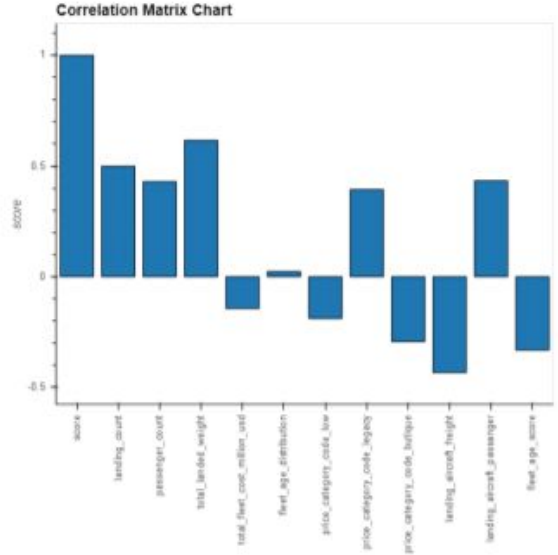
Dropped columns that were not useful and merged into one to prepare for the ML model

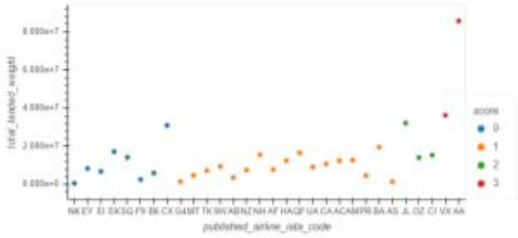
Data ETL and Classification

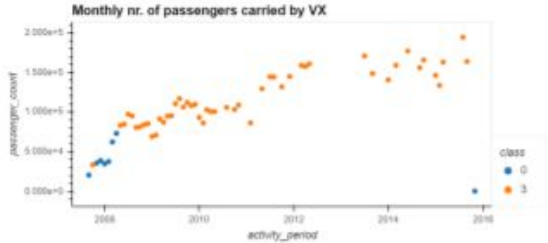
Tomasz Olewicz - Data Engineer

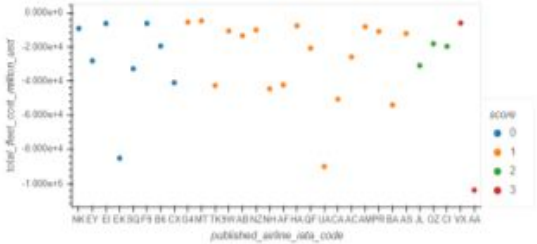
- Problem statement: Classify airlines based on their performance score
- Phase I: ETL
  1. Imported and cleaned data tables: SFO landing statistics, SFO passenger statistics, Airline Fleet
  2. Joined data files into meta table "airline_performance" and loaded to postregSQL
- Phase II: Classification and Evaluation
  1. Converted "airline_performance" to 3 Prim. Components → # of clusters from k-means
  2. Use correlation matrix to validate impact of supply data → assign rank to each group
- Phase III: Plot airline performance

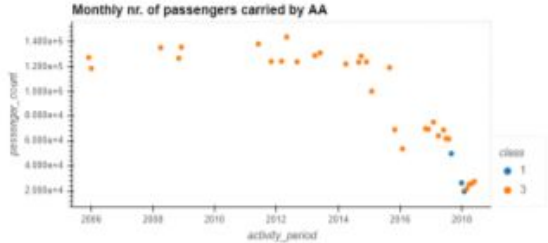


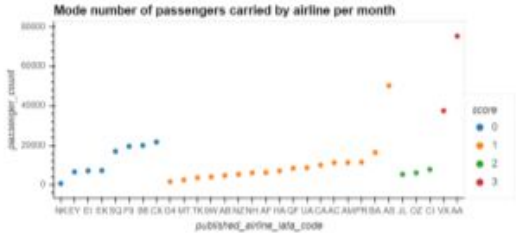

Correlation Matrix Chart
score
1
0.5
0
-0.5
score
landing_count
passenger_count
total_landed_weight
total_fleet_cost_million_usd
fleet_age_distribution
price_category_code_low
price_category_code_legacy
price_category_code_bullspot
landing_aircraft_height
landing_aircraft_passenger
fleet_age_score

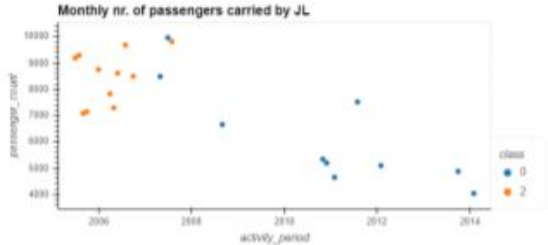

total_landed_weight
published_airline_iata_code
score
0
1
2
3

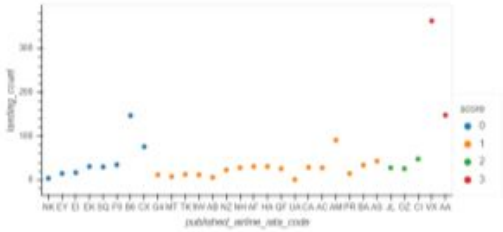

Monthly nr. of passengers carried by VX
passenger_count
activity_period
class
0
3

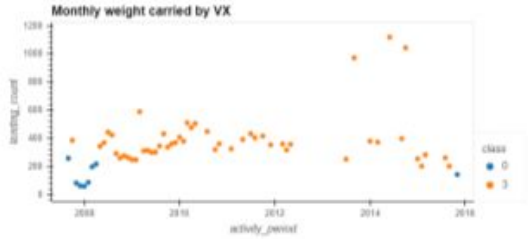

total_fleet_cost_million_usd
published_airline_iata_code
score
0
1
2
3

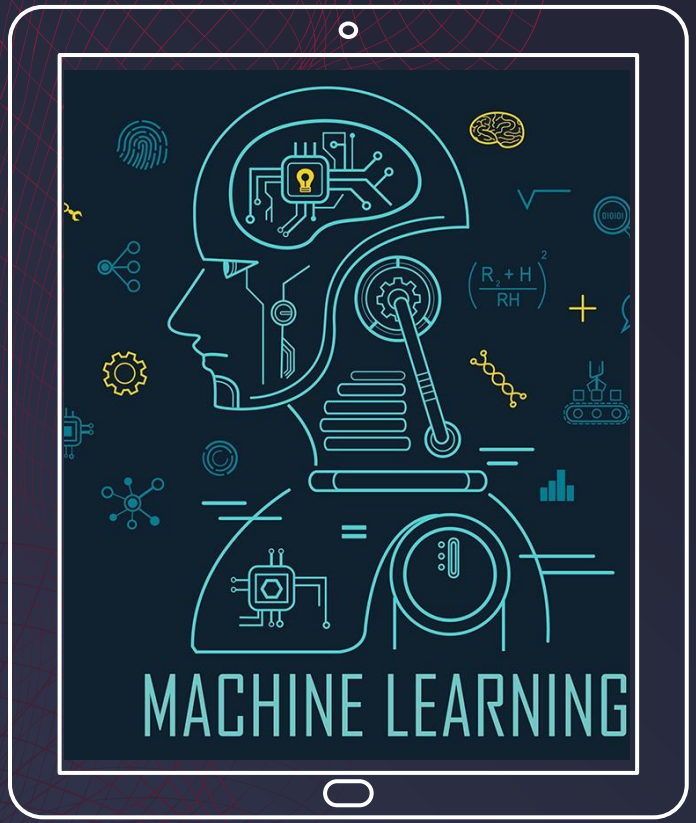

Monthly nr. of passengers carried by AA
passenger_count
activity_period
class
1
3


Mode number of passengers carried by airline per month
passenger_count
published_airline_iata_code
score
0
1
2
3


Monthly nr. of passengers carried by JL
passenger_count
activity_period
class
0
2

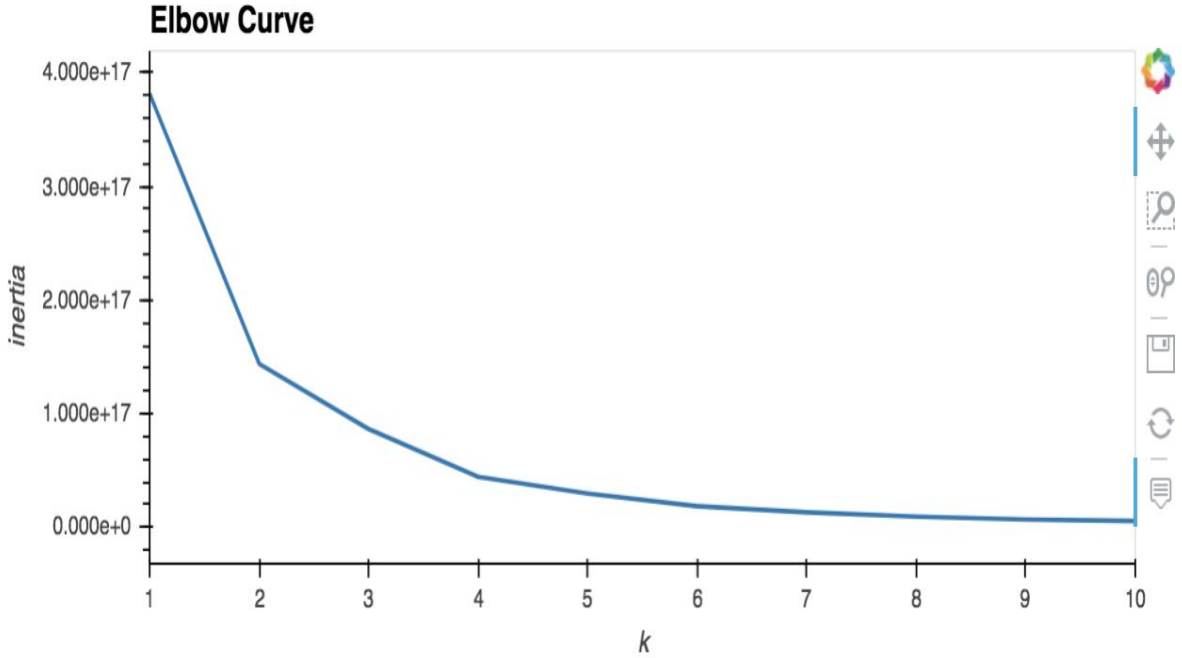

landing_count
published_airline_iata_code
score
0
1
2
3


Monthly weight carried by VX
landing_count
activity_period
class
0
3



Continued Analysis through Machine Learning

Linear Regression Model &
Support Vector Machine
using Scikit Library



Irene Kang - ML Specialist

published_airline_iata_code	Prediction	Actual
AC	0	2
QF	0	0
SQ	2	2
B6	0	0
AA	0	1
...	...	...
VX	2	2
SQ	0	0
AC	2	2
B6	1	1
VX	2	2

	pre	rec	spe	f1	geo	iba	sup
0	0.48	0.72	0.64	0.57	0.68	0.46	123
1	0.51	0.62	0.93	0.56	0.76	0.56	40
2	0.92	0.64	0.92	0.75	0.77	0.57	223
3	0.00	0.00	1.00	0.00	0.00	0.00	3
avg / total	0.73	0.66	0.83	0.67	0.73	0.53	389

Logistic Regression model accuracy: 0.625

SVM model accuracy: 0.573

Dashboard

Visualization in Tableau for our Storyboard

- Bar graphs
- Pie charts
- Clearer visual information

Time to switch screen with Alex – Data Visualization Specialist

Conclusion

- Virgin airline : top ranked with the best performance
- American airline one of the top performers ; highest cost ; liability issue due to pandemic
- With our ML models; possible analysis – public transportation, general traffic
- Further analysis is necessary on different airports in the rest of the world for better accuracy