Predicting Flight Cancellations

Serdar BOZOGLAN

MS Operations Research

Data Science Career Track Capstone Project March-18 Cohort

Springboard

github.com/serdarbozoglan

Mentor

Tuhin Sharma

Problem Definition

Total domestic flights in the first half of 2017 in top 20 airports:

~700 thousand

~10 thousand (1.28%)

Small Rate vs Big Problem

Although 1.28% is a small rate and cancellation is a rare event, causes very challenging problems for all stakeholders, especially for customers

Would not it be perfect if we predict cancellations to overcome possible problems?

Who Might Care?

Airlines

Booking Companies

travelocity
ORBITZ
ebookers
Expedia
ORBITZ FOR BUSINESS
CheapTickets
EGENCIA
HotelClub
MrJet
asia-hotels.com
Hotwire.com
ratestogo.com
trip
venere.com
艺龙 eLong
trivago
Hotels.com

Passengers

And many more...

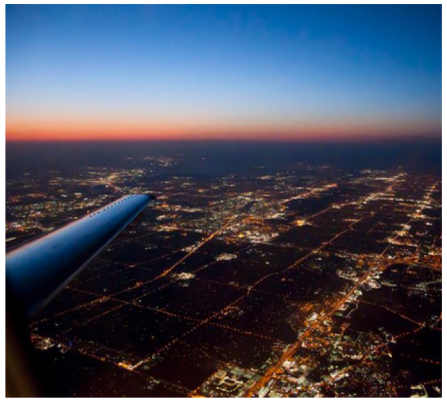
What Factors May Affect Cancellations?

Origin & Destination

Company

1. Alaska Airlines
2. United
3. Virgin America
4. JetBlue
5. American Airlines
6. Southwest
7. Delta
8. Hawaiian Airlines
9. Frontier Airlines
10. Spirit

Daytime or Night Time



Schd. Elapsed Time

Delay Minutes

Taxi Out Duration

Avg. Speed

Data Information

Data is obtained from **Department of Transportation**

Belongs to **Jan-Jun 2017**

For Top **20 Airports**

~2.7 million flights with **110** characteristic information

Assumptions

Although **weather condition** is one of the most affecting factors for flight cancellations, **we did not take weather into consideration** due to **scope of this study**

We used only top 20 airports

We used only first half of 2017 flight data to lessen computation time

Diverted flights were disregarded

Feature Engineering

Unnecessary Columns were dropped

	Unnamed: 0	Year	Quarter	Month	DayofMonth	DayOfWeek	FlightDate	Carrier	Origin	OriginCityName	OriginState	Dest	DestCityName	DestState	CRSDepTime	DepTime	DepDelay	DepDelayMinutes	DepDel15	DepartureDelayGroups	DepTimeBlk	TaxiOut	WheelsOff	WheelsOn	TaxiIn
0	0	2017	1	1	17	2	2017-01-17	AA	CLT	Charlotte, NC	NC	PHX	Phoenix, AZ	AZ	1619	1616.0	-3.0	0.0	0.0	-1.0	1600-1659	17.0	1633.0	1837.0	5.0
1	1	2017	1	1	18	3	2017-01-18	AA	CLT	Charlotte, NC	NC	PHX	Phoenix, AZ	AZ	1619	1614.0	-5.0	0.0	0.0	-1.0	1600-1659	13.0	1627.0	1815.0	6.0
2	2	2017	1	1	19	4	2017-01-19	AA	CLT	Charlotte, NC	NC	PHX	Phoenix, AZ	AZ	1619	1611.0	-8.0	0.0	0.0	-1.0	1600-1659	17.0	1628.0	1824.0	2.0

CRSArrTime	ArrTime	ArrDelay	ArrDelayMinutes	ArrDel15	ArrivalDelayGroups	ArrTimeBlk	Cancelled	CancellationCode	CRSElapsedTime	ActualElapsedTime	AirTime	Distance	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay	FirstDepTime	TotalAddGTime	LongestAddGTime	Company
1856	1842.0	-14.0	0.0	0.0	-1.0	1800-1859	0.0	NaN	277.0	266.0	244.0	1773.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	American Airlines Inc.: AA
1856	1821.0	-35.0	0.0	0.0	-2.0	1800-1859	0.0	NaN	277.0	247.0	228.0	1773.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	American Airlines Inc.: AA
1856	1826.0	-30.0	0.0	0.0	-2.0	1800-1859	0.0	NaN	277.0	255.0	236.0	1773.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	American Airlines Inc.: AA

Target feature

Feature Engineering

- **Outliers** were dropped
- The columns with more than **50% missing values** were dropped
- **Mean** method is used to fill missing values
- **New Features** are added
- Dataset **normalized**
- **Categorical** values converted into numeric values
- **Correlated** features were dropped

Data Exploration Analysis

Cancellation Rate

We split our data into train and test split with the ratio of 80% and 20% respectively.

Cancellation rate is 1.28%

Data is imbalanced

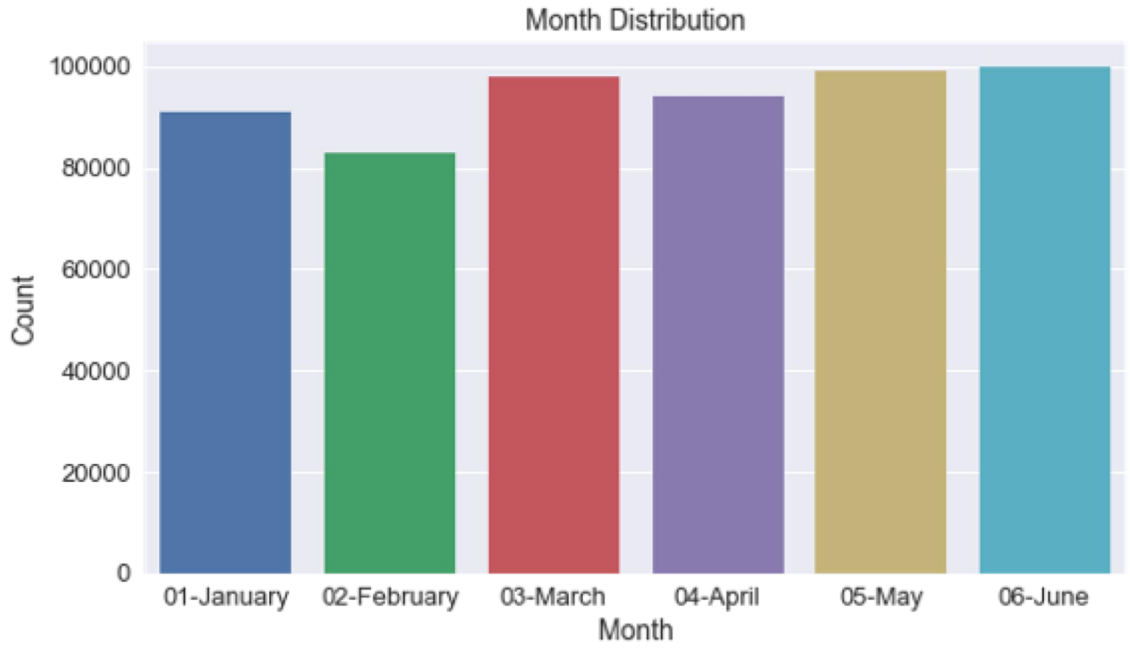
0 ➔ Non-cancelled flights
1 ➔ Cancelled flights

Data Exploration Analysis

Flight Operations Over Months

The graphic shows flight numbers for each month

June has the maximum flight operation and February has the least one.

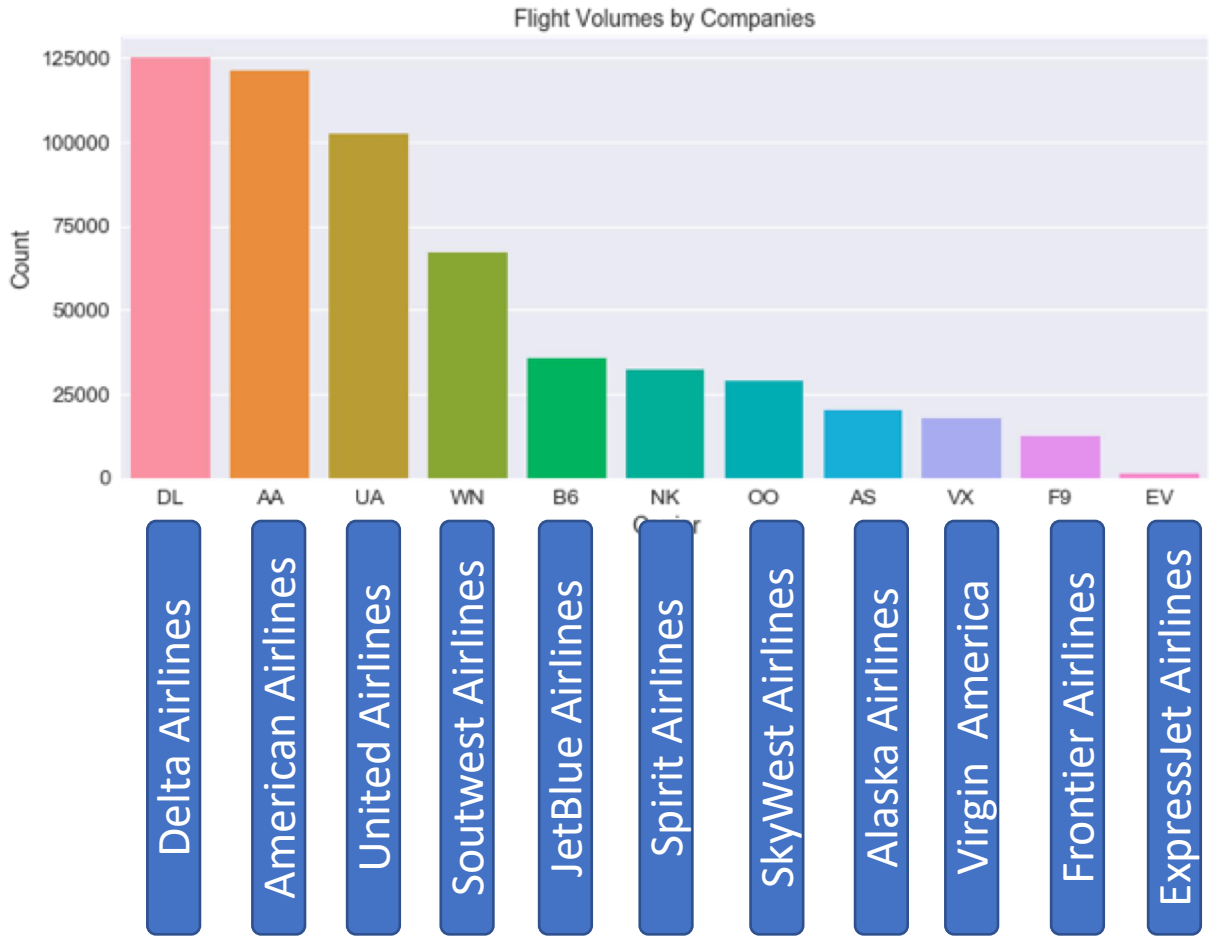


Data Exploration Analysis

Flight Volumes by Companies

The graphic shows flight numbers for each Airline company

Delta Airlines has the maximum flight operation while ExpressJet has the minimum flight operation

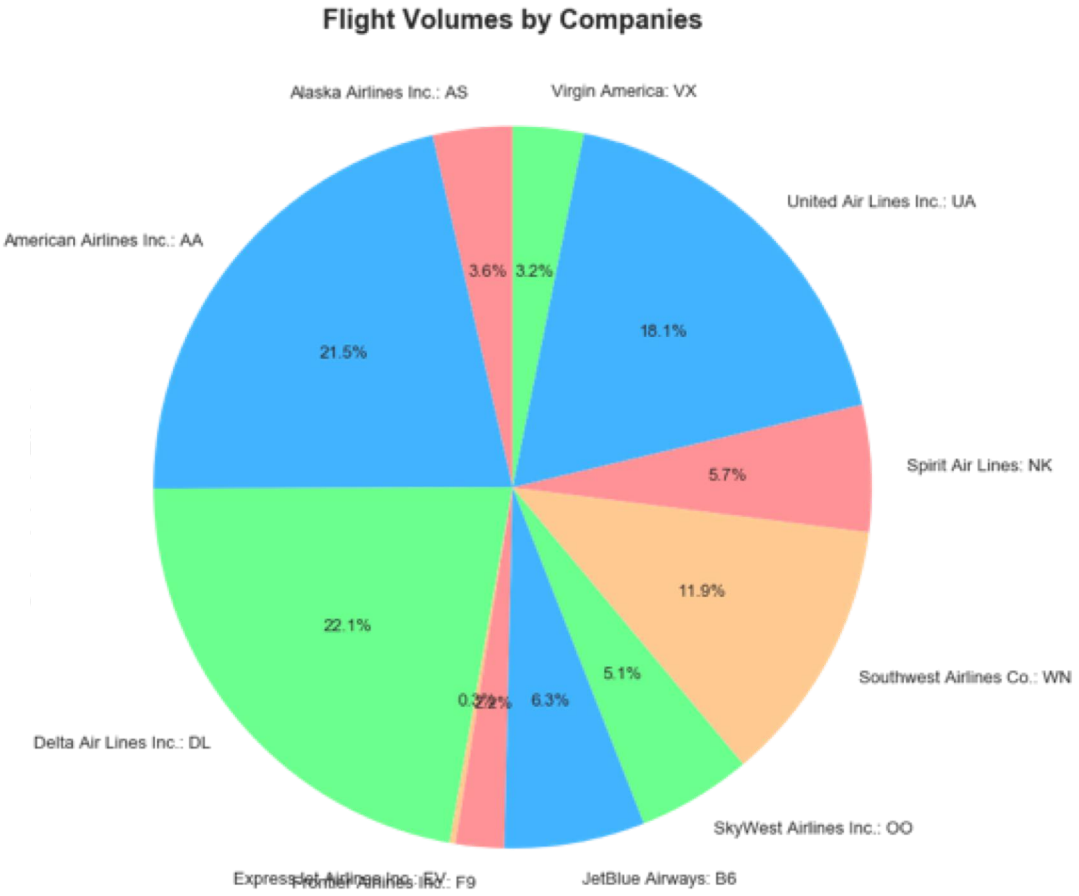


Data Exploration Analysis

Flight Volumes by Companies

The graphic shows **flight numbers** for each Airline company

Delta Airlines has the maximum flight operation while **ExpressJet** has the minimum flight operation

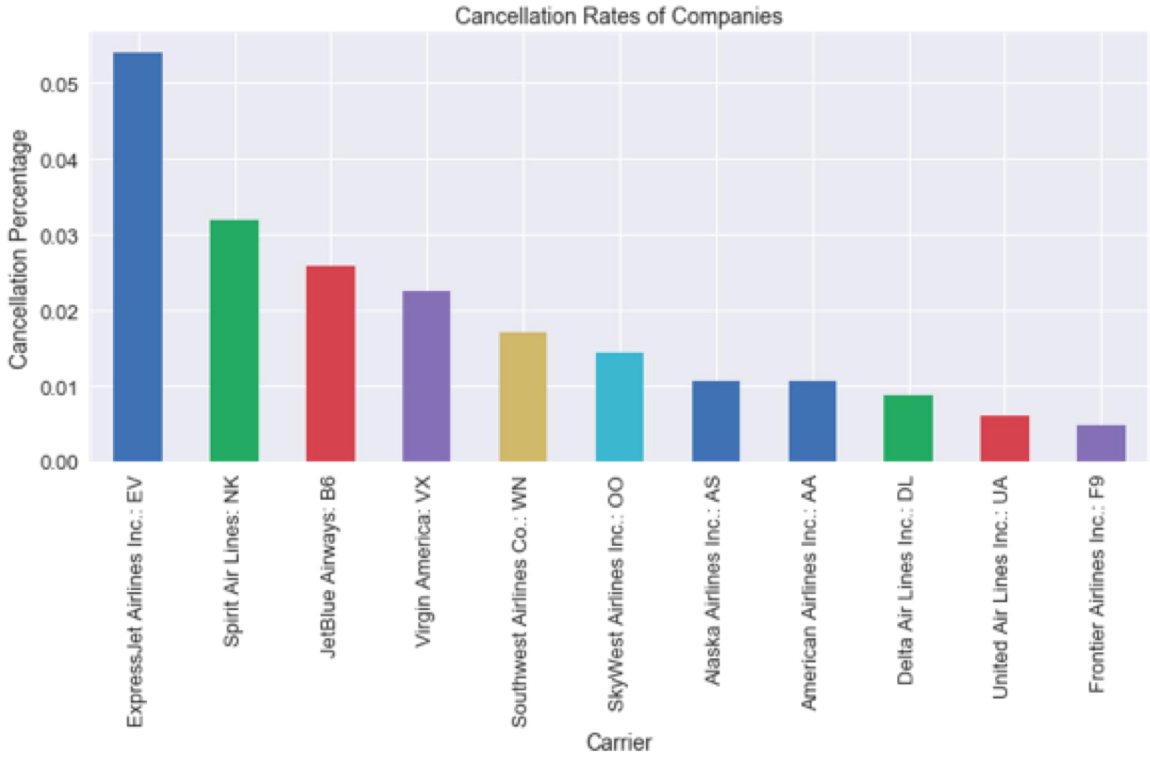


Data Exploration Analysis

Cancellation Rates by Companies

The graphic shows cancellation rates for each Airline company

Express Airlines has the maximum flight cancellation rate while Frontier Airlines has the minimum.

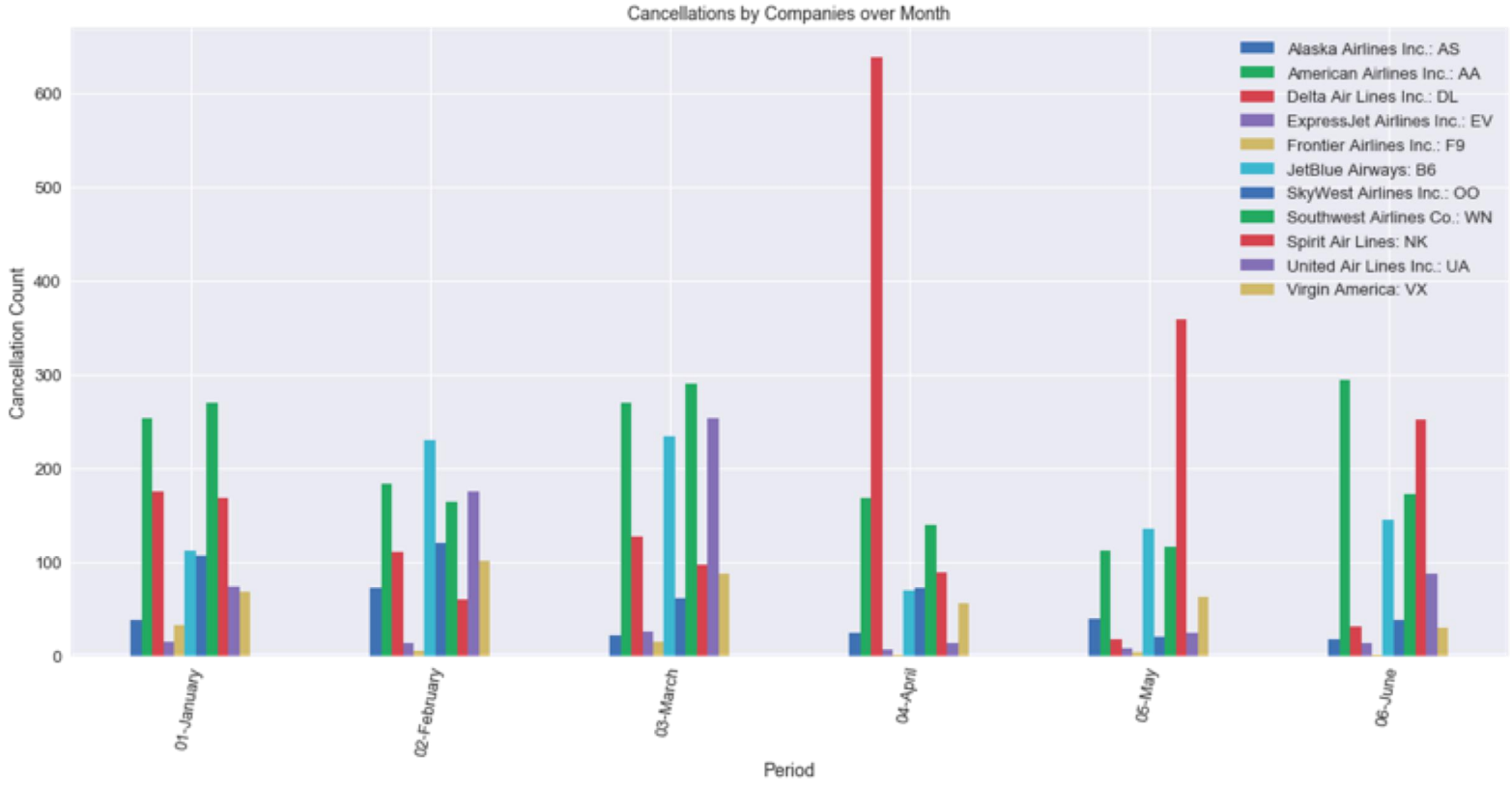


Data Exploration Analysis

Cancellation by Companies over Month

The graphic shows cancellation numbers of Airline companies over months

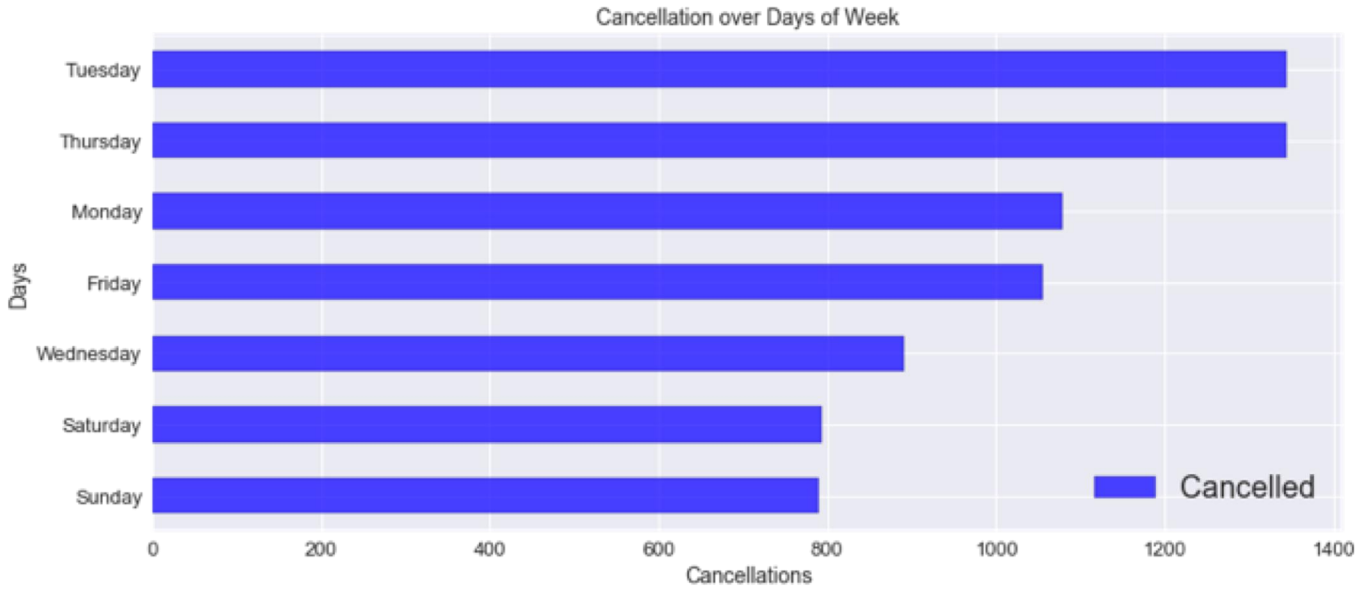
Most of the cancellations happened in March and April, especially Delta Airlines has many cancellations in April



Data Exploration Analysis

Cancellation Rates over Day of Weeks

The graphic shows cancellation numbers for each day of week



Weekdays have more cancellations than weekends

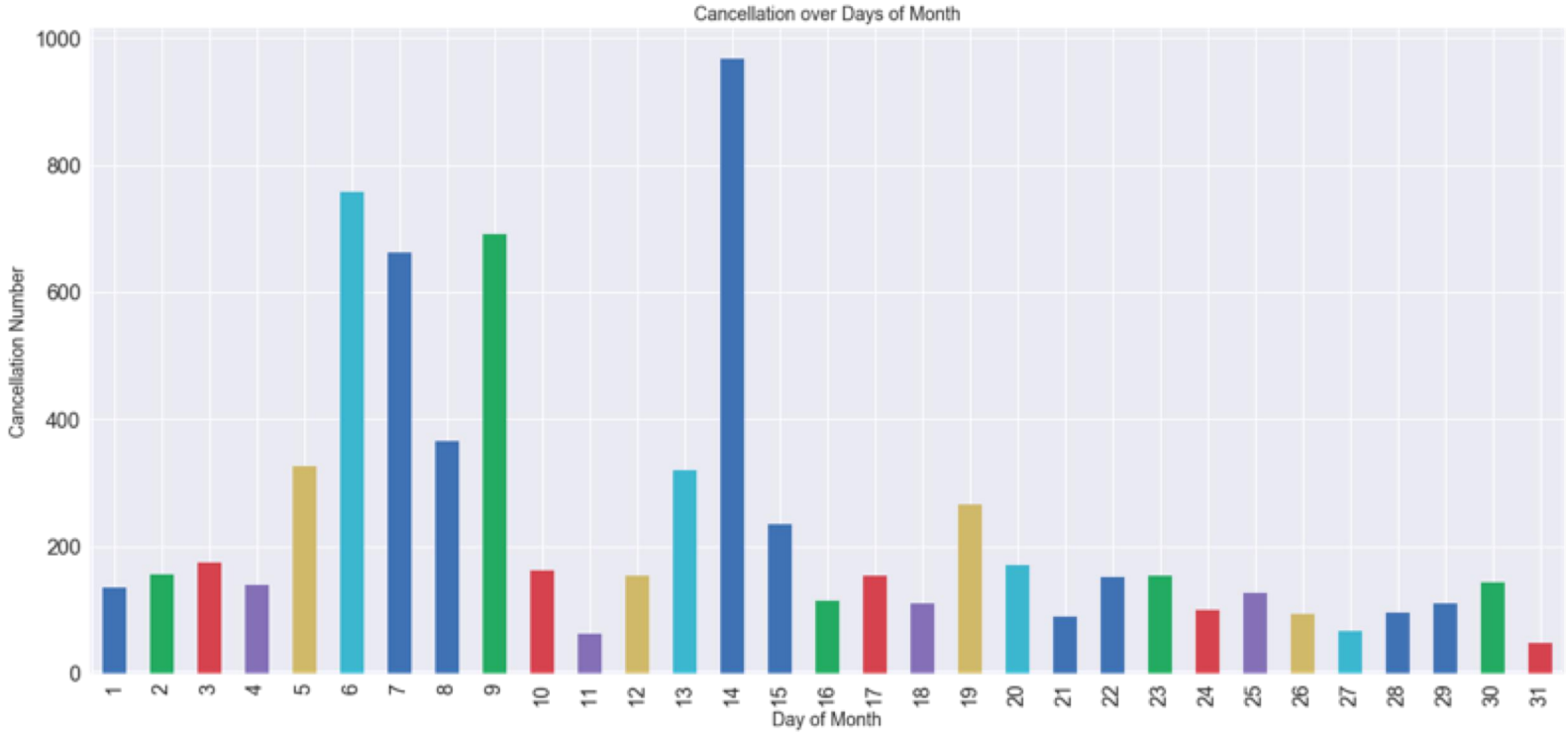
We added [PartOfWeek] new feature to our data set. We categorized the flights as weekday and weekend flights.

Data Exploration Analysis

Cancellation Rates over Day of Month

The graphic shows cancellation numbers for each day of month

First half of months have more cancellations than latter one.



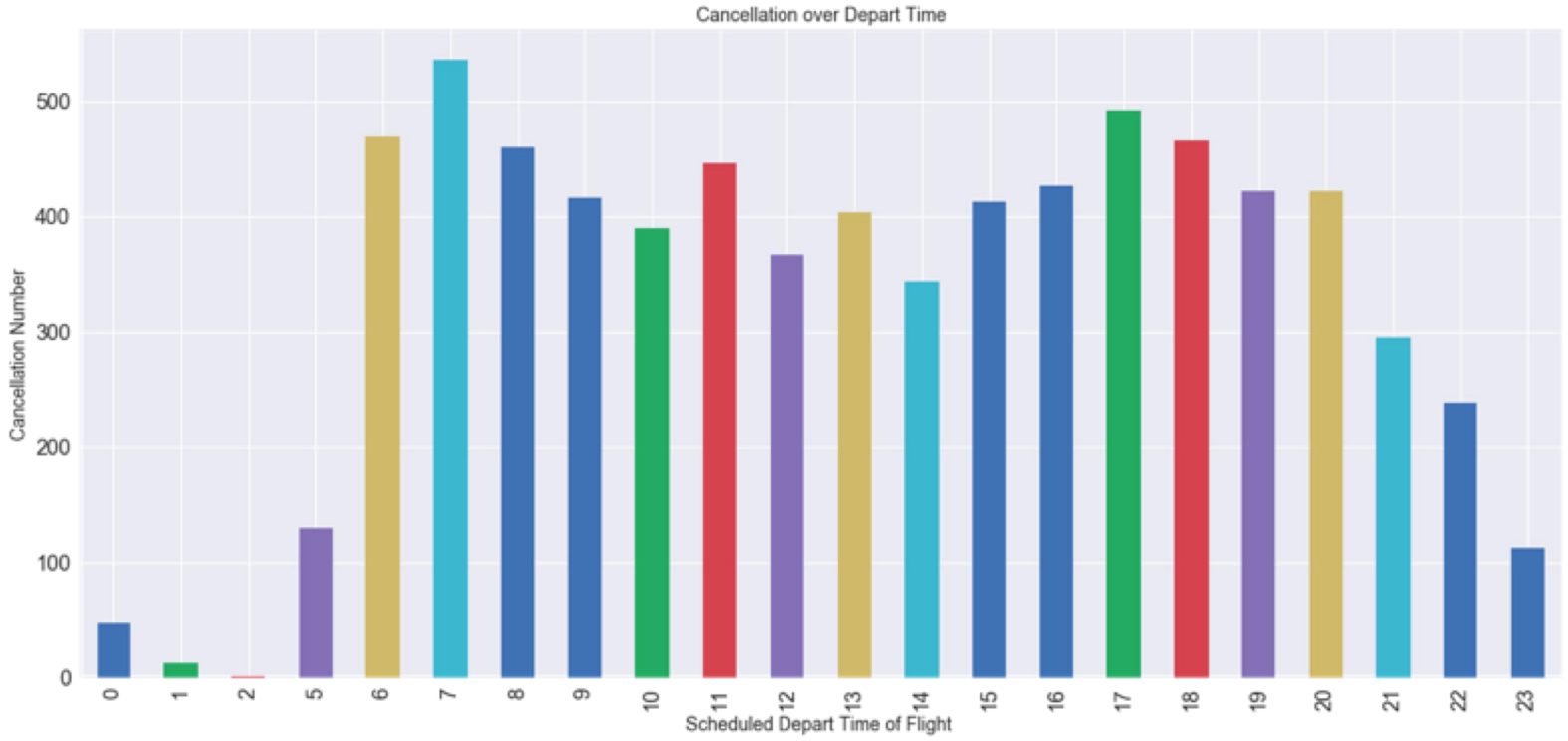
We added [PartOfMonth] new feature to our data set. We categorized the flights as first half and second half flights of month.

Data Exploration Analysis

Cancellation Rates over Depart Hour

The graphic shows cancellation numbers for each hour

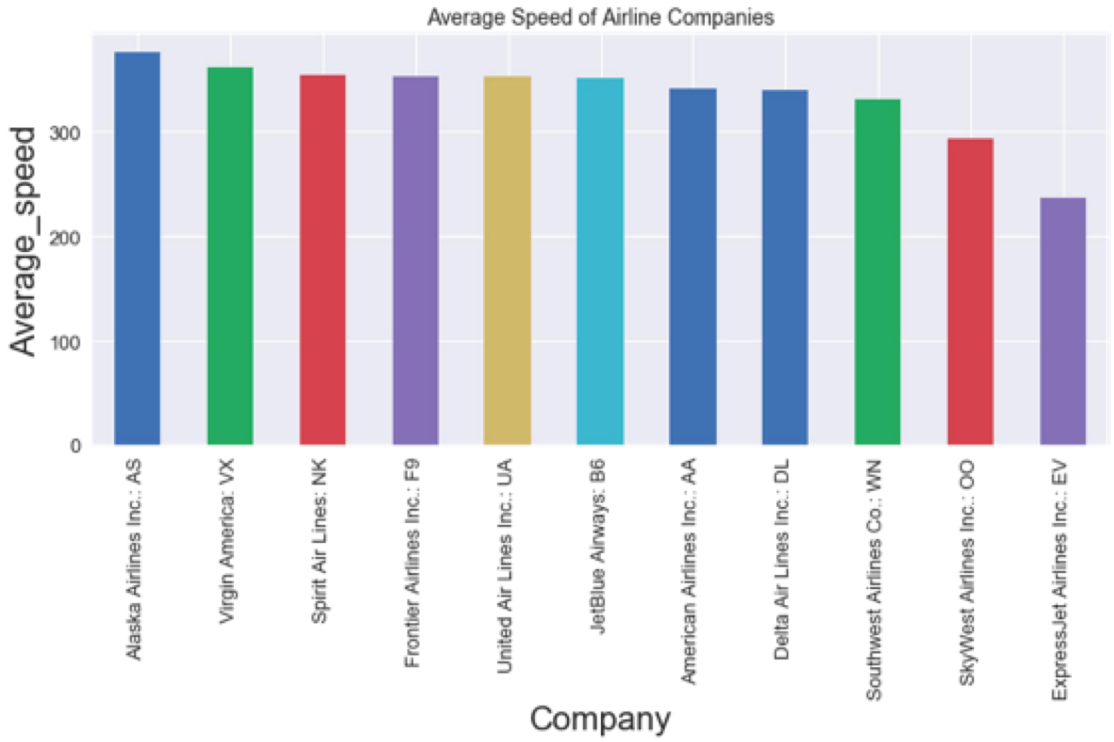
Most cancellations occurred between 6am to 9 pm (21).



We added [PartOfDay] new feature to our data set. We categorized the flights as 6am-9pm, 10pm-12 am and 1 am-5 am flights.

Data Exploration Analysis
Average Speed of Airline Companies

The graphic shows average speed for each Airline Companies



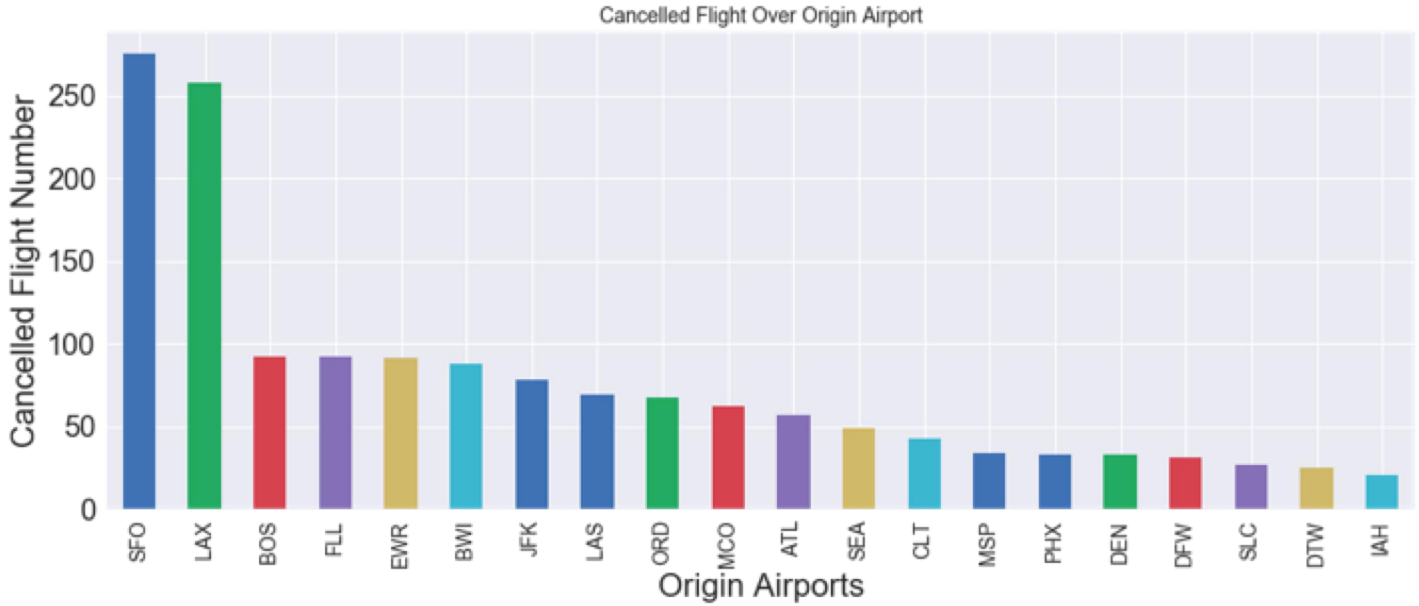
Alaska Airline is the fastest company on the other hand ExpressJet is the slowest one.

We added [AverageSpeed] new feature to our data set. We calculated average speed of companies.

Data Exploration Analysis

Cancelled Flights over Origin Airports

The graphic shows total cancelled flight numbers for each Origin Airport



ATL	Atlanta, GA
BOS	Boston, MA
BWI	Baltimore, MD
CLT	Charlotte, NC
DEN	Denver, CO
DFW	Dallas/Fort Worth, TX
DTW	Detroit, MI
EWR	Newark, NJ
IAH	Houston, TX
JFK	New York, NY
LAS	Las Vegas, NV
LAX	Los Angeles, CA
LGA	New York, NY
MCO	Orlando, FL
MSP	Minneapolis, MN
ORD	Chicago, IL
PHX	Phoenix, AZ
SEA	Seattle, WA
SFO	San Francisco, CA
SLC	Salt Lake City, UT

California (Los Angeles and San Francisco) has the most cancellations

Machine Learning Models

Overview

Type : Supervised Learning

Binary Classification (Cancelled flights shown as 1)

Tools : Scikit Learn

Models

- Logistic Regression
- k-Nearest Neighbors
- Random Forest
- Adaboost
- Gradient Boost

Machine Learning Models

General Steps

Data pre-processing :

Data splitting into training and test sets (80%-20%)

Label encoding

Scaling

Cross validation (CV) for hyperparameter tuning:

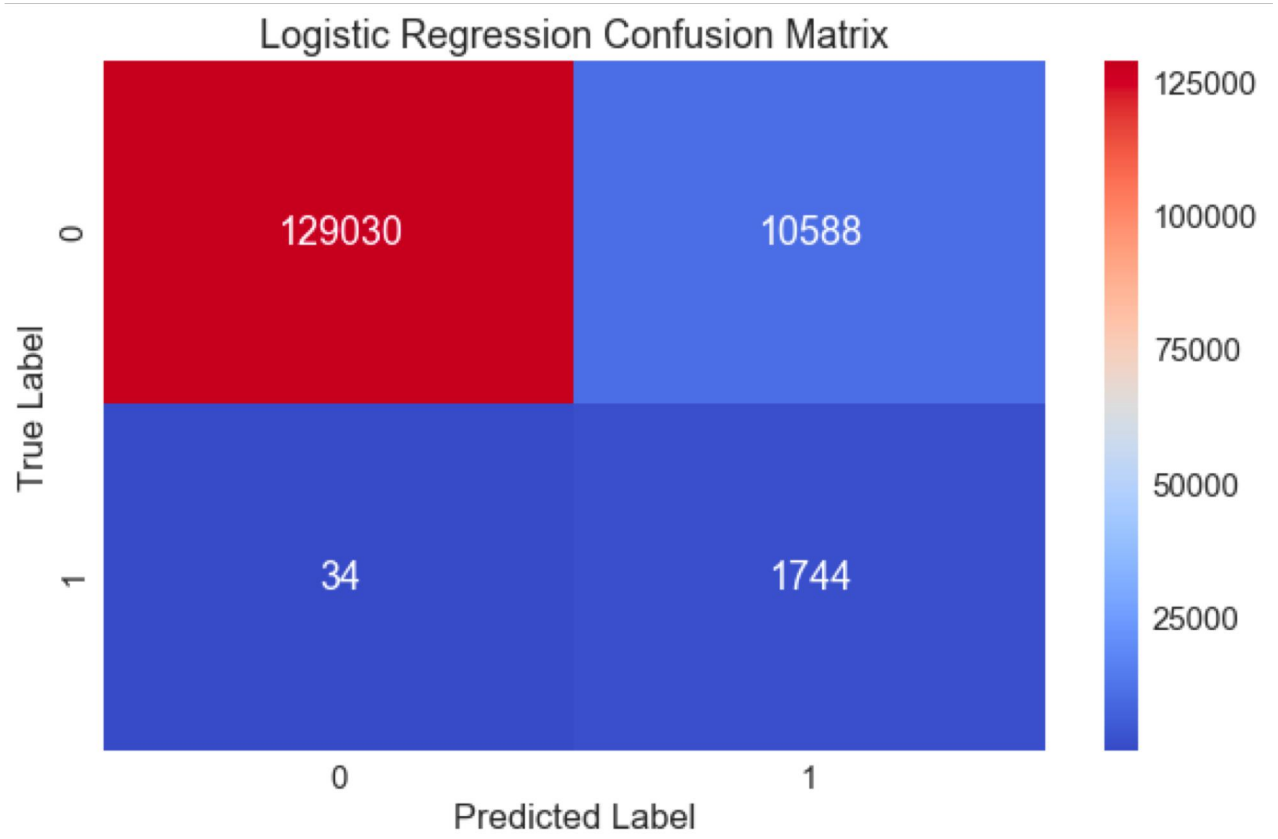
5 fold cv

Evaluation metric: Area Under ROC Curve

Machine Learning Models

Logistic Regression

Area Under ROC Curve: 0.9614

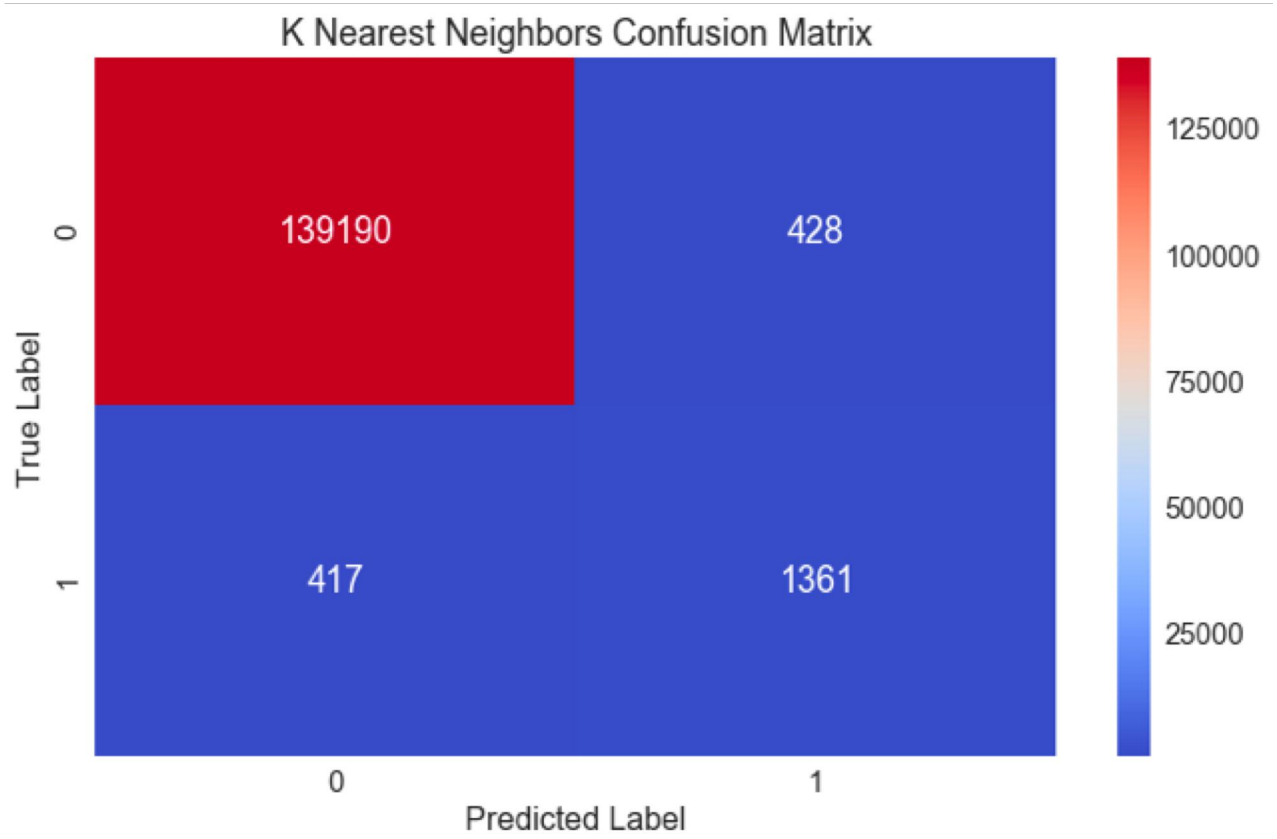


Logistic Regression is not working very well. It misclassified more than 10 thousand non-canceleld flights as cancelled. So, the False Positive Rate is incredibly high.

Machine Learning Models

k-Nearest Neighbors

Area Under ROC Curve: 0.9704

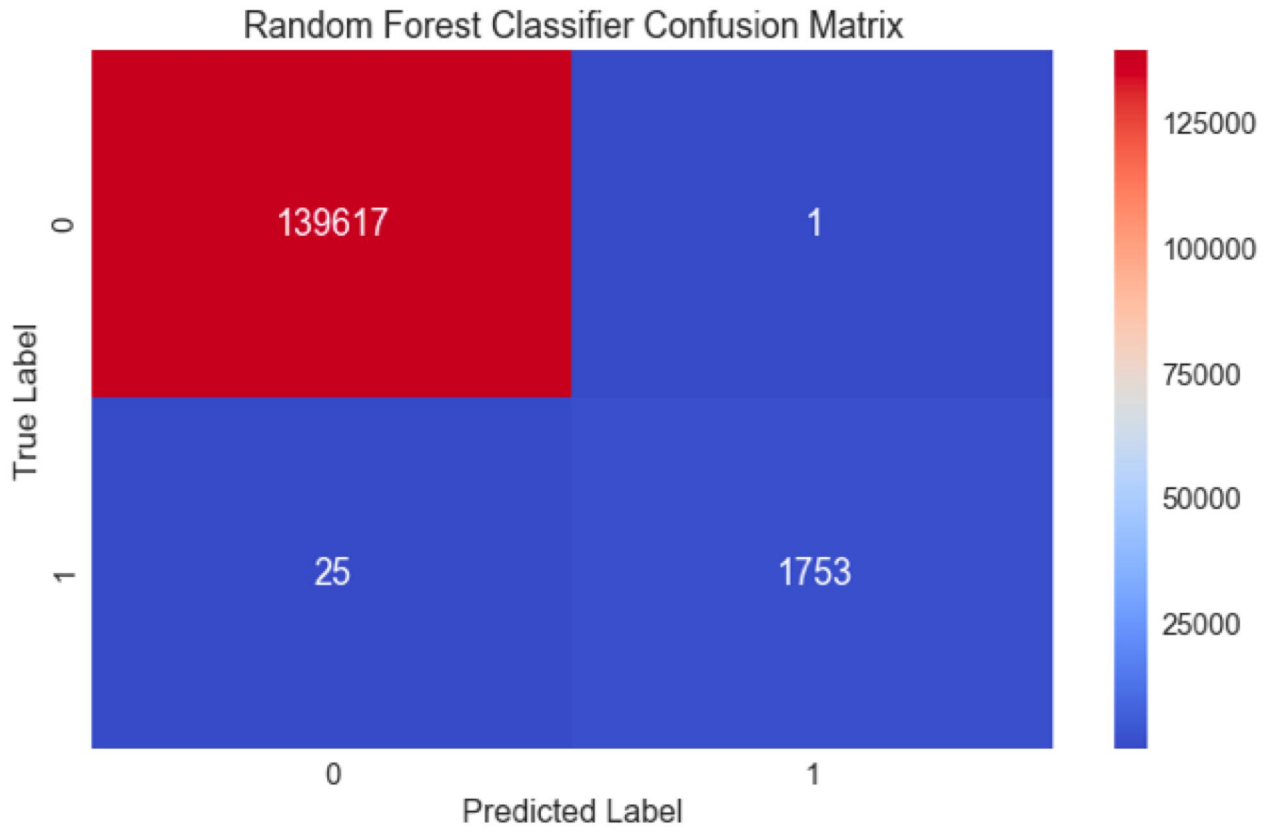


K-NN is working better than Logistic Regression. It misclassified almost 1/3 cancelled flights as non-cancelled, so Recall(sensitivity) rate is just moderate.

Machine Learning Models

Random Forest

Area Under ROC Curve: 0.9987

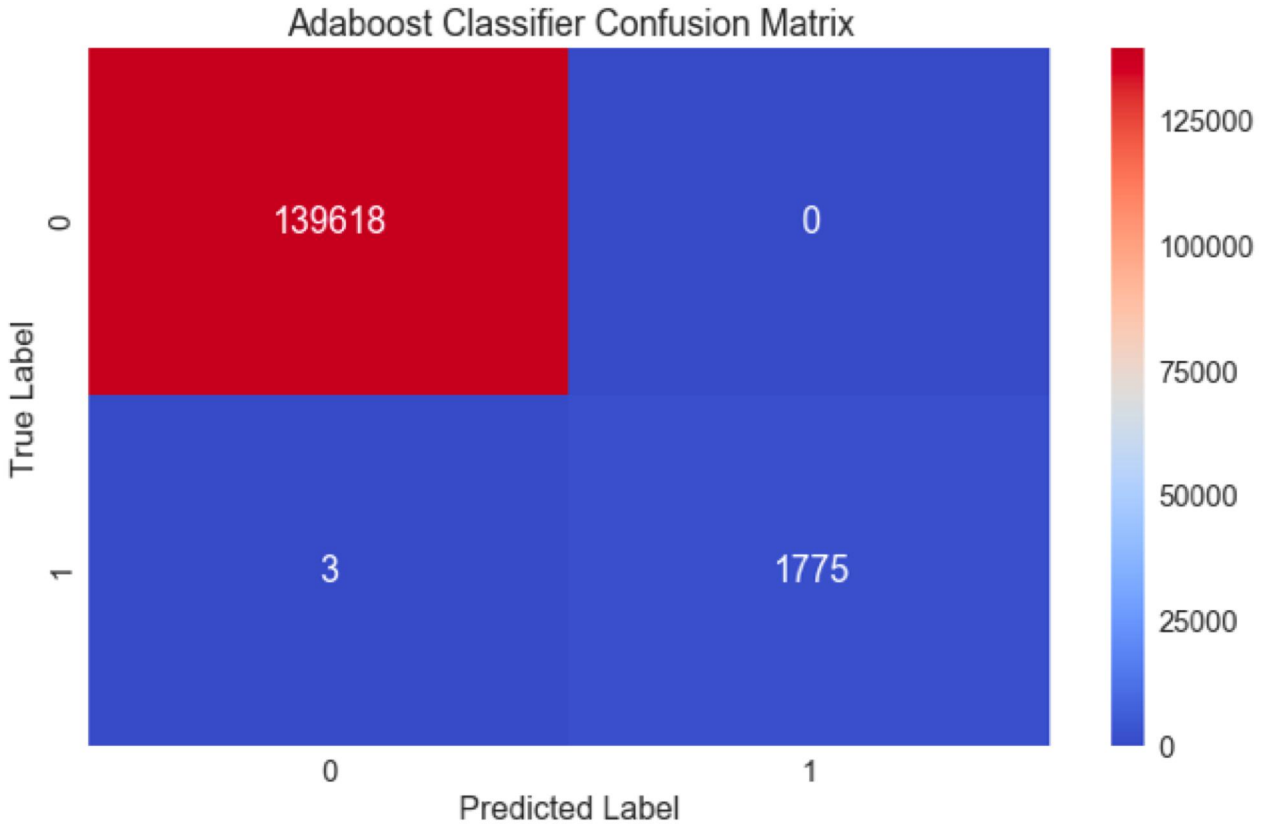


Random Forest Algorithm is working pretty well. It only missed single non-cancelled flight and 25 cancelled flights.

Machine Learning Models

Adaboost

Area Under ROC Curve: 0.9990

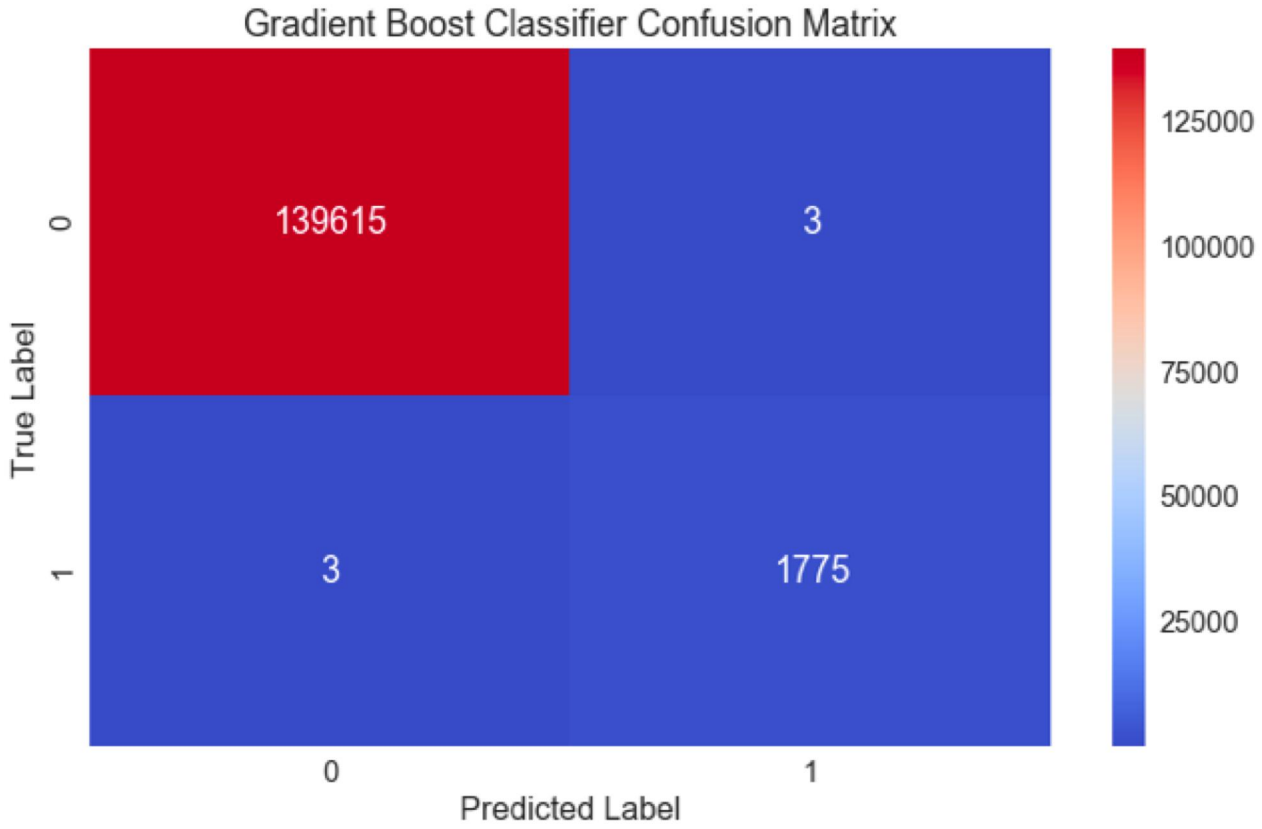


Adaboost is really performing almost perfectly. It did not misclassify any non-cancelled namely False Positive Rate is 0 which is a very good result. On the hand it only missed 3 out of 1778 cancelled flights

Machine Learning Models

Gradient Boost

Area Under ROC Curve: 0.9983



Gradient Boost is also working very well, and pretty close to Adaboost. It only misclassified 3 cancelled flights as non-cancelled and 3 non-cancelled flights as cancelled. Its sensitivity and precision are also very good. This is the second-best algorithm for this classification problem.

Machine Learning Models

Model Comparison

Model	AUC	Sensitivity(Recall)	Precision	Log Loss
Logistic Regression	0.961	0.98	0.14	0.32
K-Nearest Neighbor	0.970	0.77	0.76	0.05
Random Forest	0.998	0.99	1.0	0.002
Adaboost	0.999	1.0	1.0	0.61
Gradient Boost	0.998	1.0	1.0	0.02

Logistic Regression is the worst and Adaboost is the best classifier

Future Study

In this study, we have not focus on feature selection and model optimization. As a future study, we will concentrate on the model optimization by feature selection and hyperparameter tuning.

Conclusion

In our study, we have used all necessary features (all the one left after dropped features) in our models. According to our models, Adaboost classifier produces the outstanding scores so it is the best algorithm for our study. Gradient Boost also showed a very good performance and its evaluation metrics are very close to Adaboost so it is the second best one. On the other hand, Logistic Regression output the poorest performance with many misclassifications. We can say Random Forest is also working well. Its results are satisfactory. As for K-NN, it is better than Logistic Regression but the slowest algorithm (in terms of computation time) due to its nature.

Thank You

- Serdar BOZOGLAN
- Email: zserdarb@hotmail.com
- https://www.linkedin.com/in/serdarbozoglan/
- https://github.com/serdarbozoglan