REPORT-CAPSONE PROJECT-1

1. PROBLEM DEFINITION

Binary classification on Predicting Flight Cancellation

a. Client:

(1) Travelers:

Flight cancellations have a huge impact on whoever will have a flight trip, so travelers definitely can care cancellations.

(2) Travel Planners:

Some organization companies like tourism agency make plan for their organizations. For instance, they sell cultural/historical or holiday travel package. If their flight cancelled, then they may have encounter big problems. In this context, they also care about flight cancellation prediction.

(3) Online Booking Companies:

The companies like Booking.com, Kayak.com and Skyscanner.com are the top online flight ticket selling companies. Cancelled flights will definitely affect their business negatively. If they know in advance/ predict cancellations, they can inform their customers to take precaution against the problem which may stem from flight cancellation.

(4) Airline Companies:

Airline companies may suffer from cancellation very much. If they have relatively higher cancellation rates they may have a customer churn problem. In order to avoid this problem, they also care it.

(5) Hotels:

Even hotels, especially located nearby airports (destination) may be affected by cancelled flights. They may show interest on predicting cancelled flights.

b. Data Set:

(1) The data has been obtained from the "Bureau of Transportation Statistics" (https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time.)

(2) Data includes 110 features and 2798209 data points/observations.

(3) The data seems raw. Has too many missing values, outliers.

(4) This data includes observations from first half of 2017.

(5) Target feature is ['Cancelled'] which takes binary value, 0 for Non-cancelled flights and 1 for Cancelled flights.

2. DATA WRANGLING

Features names and their explanations presented below

Year	Year
Quarter	Quarter (1-4)
Month	Month
DayofMonth	Day of Month
DayOfWeek	Day of Week
FlightDate	Flight Date (yyyymmdd)
UniqueCarrier	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
AirlineID	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
Carrier	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.
TailNum	Tail Number
FlightNum	Flight Number
OriginAirportID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
OriginAirportSeqID	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
OriginCityMarketID	Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
Origin	Origin Airport
OriginCityName	Origin Airport, City Name
OriginState	Origin Airport, State Code
OriginStateFips	Origin Airport, State Fips
OriginStateName	Origin Airport, State Name

OriginWac	Origin Airport, World Area Code
DestAirportID	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused.
DestAirportSeqID	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.
DestCityMarketID	Destination Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market.
Dest	Destination Airport
DestCityName	Destination Airport, City Name
DestState	Destination Airport, State Code
DestStateFips	Destination Airport, State Fips
DestStateName	Destination Airport, State Name
DestWac	Destination Airport, World Area Code
CRSDepTime	CRS Departure Time (local time: hhmm)
DepTime	Actual Departure Time (local time: hhmm)
DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DepDelayMinutes	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
DepDel15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
DepartureDelayGroups	Departure Delay intervals, every (15 minutes from <-15 to >180)
DepTimeBlk	CRS Departure Time Block, Hourly Intervals
TaxiOut	Taxi Out Time, in Minutes
WheelsOff	Wheels Off Time (local time: hhmm)
WheelsOn	Wheels On Time (local time: hhmm)
TaxiIn	Taxi In Time, in Minutes
CRSArrTime	CRS Arrival Time (local time: hhmm)
ArrTime	Actual Arrival Time (local time: hhmm)
ArrDelay	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ArrDelayMinutes	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ArrDel15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
ArrivalDelayGroups	Arrival Delay intervals, every (15-minutes from <-15 to >180)
ArrTimeBlk	CRS Arrival Time Block, Hourly Intervals
Cancelled	Cancelled Flight Indicator (1=Yes)

CancellationCode	Specifies The Reason For Cancellation
Diverted	Diverted Flight Indicator (1=Yes)
CRSElapsedTime	CRS Elapsed Time of Flight, in Minutes
ActualElapsedTime	Elapsed Time of Flight, in Minutes
AirTime	Flight Time, in Minutes
Flights	Number of Flights
Distance	Distance between airports (miles)
DistanceGroup	Distance Intervals, every 250 Miles, for Flight Segment
CarrierDelay	Carrier Delay, in Minutes
WeatherDelay	Weather Delay, in Minutes
NASDelay	National Air System Delay, in Minutes
SecurityDelay	Security Delay, in Minutes
LateAircraftDelay	Late Aircraft Delay, in Minutes
FirstDepTime	First Gate Departure Time at Origin Airport
TotalAddGTime	Total Ground Time Away from Gate for Gate Return or Cancelled Flight
LongestAddGTime	Longest Time Away from Gate for Gate Return or Cancelled Flight
DivAirportLandings	Number of Diverted Airport Landings
DivReachedDest	Diverted Flight Reaching Scheduled Destination Indicator (1=Yes)
DivActualElapsedTime	Elapsed Time of Diverted Flight Reaching Scheduled Destination, in Minutes. The ActualElapsedTime column remains NULL for all diverted flights.
DivArrDelay	Difference in minutes between scheduled and actual arrival time for a diverted flight reaching scheduled destination. The ArrDelay column remains NULL for all diverted flights.
DivDistance	Distance between scheduled destination and final diverted airport (miles). Value will be 0 for diverted flight reaching scheduled destination.
Div1Airport	Diverted Airport Code1
Div1AirportID	Airport ID of Diverted Airport 1. Airport ID is a Unique Key for an Airport
Div1AirportSeqID	Airport Sequence ID of Diverted Airport 1. Unique Key for Time Specific Information for an Airport
Div1WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code1
Div1TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code1
Div1LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code1
Div1WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code1
Div1TailNum	Aircraft Tail Number for Diverted Airport Code1
Div2Airport	Diverted Airport Code2
Div2AirportID	Airport ID of Diverted Airport 2. Airport ID is a Unique Key for an Airport

Div2AirportSeqID	Airport Sequence ID of Diverted Airport 2. Unique Key for Time Specific Information for an Airport
Div2WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code2
Div2TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code2
Div2LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code2
Div2WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code2
Div2TailNum	Aircraft Tail Number for Diverted Airport Code2
Div3Airport	Diverted Airport Code3
Div3AirportID	Airport ID of Diverted Airport 3. Airport ID is a Unique Key for an Airport
Div3AirportSeqID	Airport Sequence ID of Diverted Airport 3. Unique Key for Time Specific Information for an Airport
Div3WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code3
Div3TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code3
Div3LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code3
Div3WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code3
Div3TailNum	Aircraft Tail Number for Diverted Airport Code3
Div4Airport	Diverted Airport Code4
Div4AirportID	Airport ID of Diverted Airport 4. Airport ID is a Unique Key for an Airport
Div4AirportSeqID	Airport Sequence ID of Diverted Airport 4. Unique Key for Time Specific Information for an Airport
Div4WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code4
Div4TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code4
Div4LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code4
Div4WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code4
Div4TailNum	Aircraft Tail Number for Diverted Airport Code4
Div5Airport	Diverted Airport Code5
Div5AirportID	Airport ID of Diverted Airport 5. Airport ID is a Unique Key for an Airport
Div5AirportSeqID	Airport Sequence ID of Diverted Airport 5. Unique Key for Time Specific Information for an Airport
Div5WheelsOn	Wheels On Time (local time: hhmm) at Diverted Airport Code5
Div5TotalGTime	Total Ground Time Away from Gate at Diverted Airport Code5
Div5LongestGTime	Longest Ground Time Away from Gate at Diverted Airport Code5
Div5WheelsOff	Wheels Off Time (local time: hhmm) at Diverted Airport Code5
Div5TailNum	Aircraft Tail Number for Diverted Airport Code5

REPORT-CAPSONE PROJECT-1

	Year	Quarter	Month	DayofMonth	DayOfWeek	FlightDate	UniqueCarrier	AirlineID	Carrier	TailNum	FlightNum	OriginAirportID	OriginAirportSeqID	OriginCityMarketID	Origin
0	2017	1	1	17	2	2017-01-17	AA	19805	AA	N583AA	494	11057	1105703	31057	CLT
1	2017	1	1	18	3	2017-01-18	AA	19805	AA	N544AA	494	11057	1105703	31057	CLT
2	2017	1	1	19	4	2017-01-19	AA	19805	AA	N553AA	494	11057	1105703	31057	CLT
3	2017	1	1	20	5	2017-01-20	AA	19805	AA	N191AA	494	11057	1105703	31057	CLT
4	2017	1	1	21	6	2017-01-21	AA	19805	AA	N170AA	494	11057	1105703	31057	CLT

OriginCityName	OriginState	OriginStateFips	OriginStateName	OriginWac	DestAirportID	DestAirportSeqID	DestCityMarketID	Dest	DestCityName	DestState
Charlotte, NC	NC	37	North Carolina	36	14107	1410702	30466	PHX	Phoenix, AZ	AZ
Charlotte, NC	NC	37	North Carolina	36	14107	1410702	30466	PHX	Phoenix, AZ	AZ
Charlotte, NC	NC	37	North Carolina	36	14107	1410702	30466	PHX	Phoenix, AZ	AZ
Charlotte, NC	NC	37	North Carolina	36	14107	1410702	30466	PHX	Phoenix, AZ	AZ
Charlotte, NC	NC	37	North Carolina	36	14107	1410702	30466	PHX	Phoenix, AZ	AZ

DestStateFips	DestStateName	DestWac	CRSDepTime	DepTime	DepDelay	DepDelayMinutes	DepDel15	DepartureDelayGroups	DepTimeBlk	TaxiOut	WheelsOff
4	Arizona	81	1619	1616.0	-3.0	0.0	0.0	-1.0	1600-1659	17.0	1633.0
4	Arizona	81	1619	1614.0	-5.0	0.0	0.0	-1.0	1600-1659	13.0	1627.0
4	Arizona	81	1619	1611.0	-8.0	0.0	0.0	-1.0	1600-1659	17.0	1628.0
4	Arizona	81	1619	1656.0	37.0	37.0	1.0	2.0	1600-1659	18.0	1714.0
4	Arizona	81	1619	1632.0	13.0	13.0	0.0	0.0	1600-1659	17.0	1649.0

DestStateFips	DestStateName	DestWac	CRSDepTime	DepTime	DepDelay	DepDelayMinutes	DepDel15	DepartureDelayGroups	DepTimeBlk	TaxiOut	WheelsOff
4	Arizona	81	1619	1616.0	-3.0	0.0	0.0	-1.0	1600-1659	17.0	1633.0
4	Arizona	81	1619	1614.0	-5.0	0.0	0.0	-1.0	1600-1659	13.0	1627.0
4	Arizona	81	1619	1611.0	-8.0	0.0	0.0	-1.0	1600-1659	17.0	1628.0
4	Arizona	81	1619	1656.0	37.0	37.0	1.0	2.0	1600-1659	18.0	1714.0
4	Arizona	81	1619	1632.0	13.0	13.0	0.0	0.0	1600-1659	17.0	1649.0

WheelsOn	TaxiIn	CRSArrTime	ArrTime	ArrDelay	ArrDelayMinutes	ArrDel15	ArrivalDelayGroups	ArrTimeBlk	Cancelled	CancellationCode	Diverted	CRSElapsedTime
1837.0	5.0	1856	1842.0	-14.0	0.0	0.0	-1.0	1800-1859	0.0	NaN	0.0	277.0
1815.0	6.0	1856	1821.0	-35.0	0.0	0.0	-2.0	1800-1859	0.0	NaN	0.0	277.0
1824.0	2.0	1856	1826.0	-30.0	0.0	0.0	-2.0	1800-1859	0.0	NaN	0.0	277.0
1926.0	3.0	1856	1929.0	33.0	33.0	1.0	2.0	1800-1859	0.0	NaN	0.0	277.0
1854.0	4.0	1856	1858.0	2.0	2.0	0.0	0.0	1800-1859	0.0	NaN	0.0	277.0

REPORT-CAPSONE PROJECT-1

ActualElapsedTime	AirTime	Flights	Distance	DistanceGroup	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay	FirstDepTime	TotalAddGTime	LongestAddGTime
266.0	244.0	1.0	1773.0	8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
247.0	228.0	1.0	1773.0	8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
255.0	236.0	1.0	1773.0	8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
273.0	252.0	1.0	1773.0	8	33.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN
266.0	245.0	1.0	1773.0	8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

DivAirportLandings	DivReachedDest	DivActualElapsedTime	DivArrDelay	DivDistance	Div1Airport	Div1AirportID	Div1AirportSeqID	Div1WheelsOn	Div1TotalGTime	Div1LongestGTime	Div1WheelsOff	Div1TailNum
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

DivAirportLandings	DivReachedDest	DivActualElapsedTime	DivArrDelay	DivDistance	Div1Airport	Div1AirportID	Div1AirportSeqID	Div1WheelsOn	Div1TotalGTime	Div1LongestGTime	Div1WheelsOff	Div1TailNum
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Div2Airport	Div2AirportID	Div2AirportSeqID	Div2WheelsOn	Div2TotalGTime	Div2LongestGTime	Div2WheelsOff	Div2TailNum	Div3Airport	Div3AirportID	Div3AirportSeqID
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Div3WheelsOn	Div3TotalGTime	Div3LongestGTime	Div3WheelsOff	Div3TailNum	Div4Airport	Div4AirportID	Div4AirportSeqID	Div4WheelsOn	Div4TotalGTime
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Div4WheelsOff	Div4TailNum	Div5Airport	Div5AirportID	Div5AirportSeqID	Div5WheelsOn	Div5TotalGTime	Div5LongestGTime	Div5WheelsOff	Div5TailNum
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

a. Dropping Irrelevant Columns at Once:

['UniqueCarrier','TailNum','FlightNum','OriginAirportID','OriginAirportSeqID','OriginCityMarketID'OriginStateFips','OriginStateName','OriginWac','DestAirportID','DestAirportSeqID',' DestCityMarketID','DestStateFips','DestStateNameDestWac','Flights','DistanceGroup','Diverted', 'DivAirportLandings','DivReachedDest','DivActualElapsedTime','DivArrDelay','DivDistance', 'Div1Airport','Div1AirportID','Div1AirportSeqID','Div1WheelsOn','Div1TotalGTime' 'Div1LongestGTime','Div1WheelsOff','Div1TailNum','Div2Airport','Div2AirportID', 'Div2AirportSeqID','Div2WheelsOn,'Div2TotalGTime','Div2LongestGTime','Div2WheelsOff', 'Div2TailNum','Div3Airport','Div3AirportID','Div3AirportSeqID,'Div3WheelsOn','Div3TotalGTime','Div3LongestGTime','Div3WheelsOff','Div3TailNum','Div4Airport','Div4AirportID', 'Div4AirportSeqID','Div4WheelsOn','Div4TotalGTime','Div4LongestGTime','Div4WheelsOff', 'Div4TailNum','Div5Airport','Div5AirportID','Div5AirportSeqID','Div5WheelsOn','Div5TotalGTime','Div5LongestGTime','Div5WheelsOff', 'Div5TailNum']

Some of those columns have redundant information such as *OriginCityMarketID* or *StateName* of Origin Airport columns and rest of them are linked with diverted flights which is not in our scope of this study. So, we decided to drop above features at the beginning of our analysis.

b. Dropping Other Columns after Analysis:

(1) 'LongestAddGTime',
(1) 'TotalAddGTime',
(2) 'FirstDepTime',
(3) 'CancellationCode',
(4) 'LateAircraftDelay',
(5) 'SecurityDelay',
(6) 'NASDelay',
(7) 'WeatherDelay',
(8) 'CarrierDelay',
(9) 'DepDel15',
(10) 'DepartureDelayGroups',
(11) 'DepDelay',
(12) 'ActualElapsedTime',
(13) 'AirTime',

(14) 'WheelsOff',
(15) 'WheelsOn',
(16) 'TaxiIn',
(17) 'ArrTime',
(18) 'ArrDelayMinutes',
(19) 'ArrDel15',
(20) 'ArrivalDelayGroups',
(21) 'CRSDepTime',
(22) 'CRSArrTime',
(23) 'Year',
(24) 'OriginCityName',
(25) 'OriginState',
(26) 'DestCityName',
(27) 'DestState',
(28) 'DepTimeBlk',
(29) 'ArrTimeBlk',
(30) 'DepDelay',
(31) 'DepTime',
(32) 'ArrDelay',
(33) 'FlightDate',
(34) 'FlightDate_Dest',
(35) 'Distance',
(36) 'Quarter',
(37) 'Month',
(38) 'DayofMonth',
(39) 'DayOfWeek',
(40) 'Flight_Hour',
(41) 'Company',
(42) 'FlightDateTime_Origin',
(43) 'FlightDateTime_Dest',
(44) 'MonthName',
(45) 'DayOfWeekName'

Those columns were dropped after analysis and after using in EDA. Those columns are also unnecessary for our analysis.

c. Fixing DateTime Format

We have **FlightDate** and **CRSDepTime (Scheduled Departure Time)**, **CRSArrTime (Scheduled Arrival Time)** columns with date and time information in our dataframe. First, we will convert FlightDate to datetime object and then we will create FlightDate_Dest because date may change when the airplane lands on the destination due to time zone difference and flight duration. Then we will create a flight DATETIME at origin (FlightDateTime_Origin) and destination (FlightDateTime_Dest)

d. Changing Data Type:

Float type of Cancelled column converted into integer.

e. Cleansing Outliers

There were some observations required unrealistic airliner speed (for instance 1000 miles/hour) to fulfill that flight. It is impossible to be real, so those observations considered as an outlier and dropped.

f. Getting Rid of Null/Missing Values

(1) Threshold to Drop:

50% missing values in a column was defined as a threshold to remove from the dataset because it does not give us information. So the features with 50% or above missing values were deleted. In this context, some features [LongestAddGTime, TotalAddGTime, FirsDepTime, CancellationCode, LateAircraftDelay, SecurityDelay, NASDelay, WeatherDelay, CarrierDelay] were dropped.

(2) Mean to Fill:

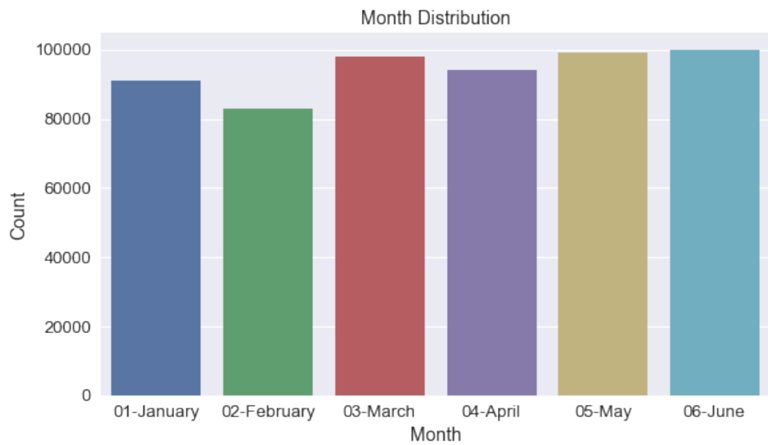
I had a difficult time to fill missing values and could not figure out how to handle. But later I evaluated that the feature of [DepDelayMinutes] and [TaxiOut] may give us good information reading predicting the cancellation of flights. Those columns had 1.5% missing values. I figure out that almost only some of the cancelled flights had those missing values but non-cancelled. In this context, mean of cancelled flights was used to impute the missing values.

3. EXPLORATORY DATA ANALYSIS (EDA)-DATA VISUALIZATION

a. Target Feature Distribution:

In the graphic, 1 represents Cancelled flights and 0 Non-cancelled. As seen fron graphic, the Cancellation rate is fairly unbalanced.

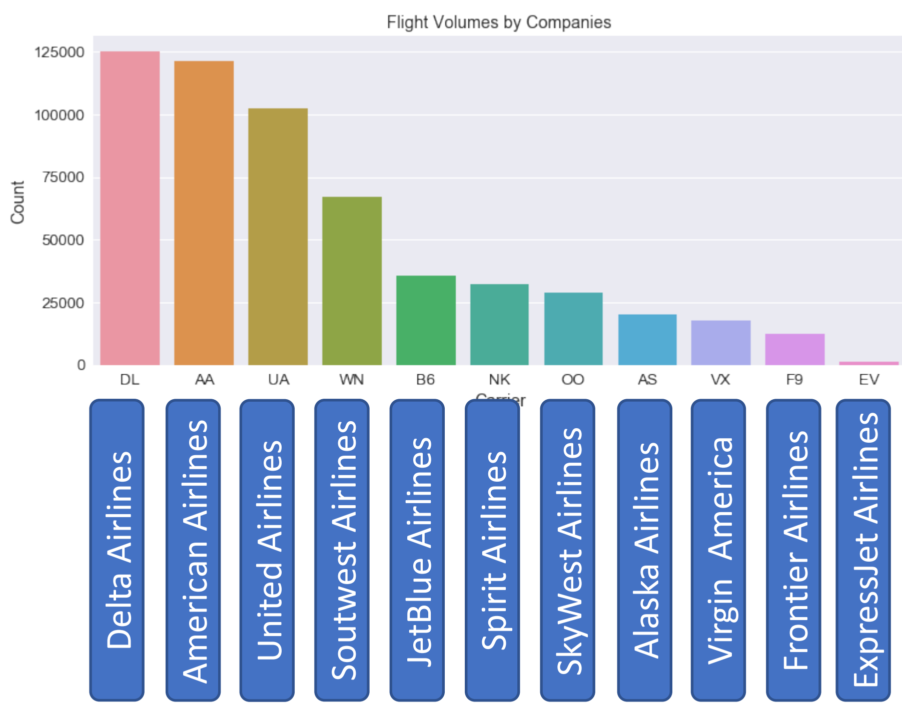
b. Flight Over Months:



The graphic shows flight numbers for each month.

The flights are almost fairly distributed among the months. As we would expect, due to be the shortest month February has the least flight number.

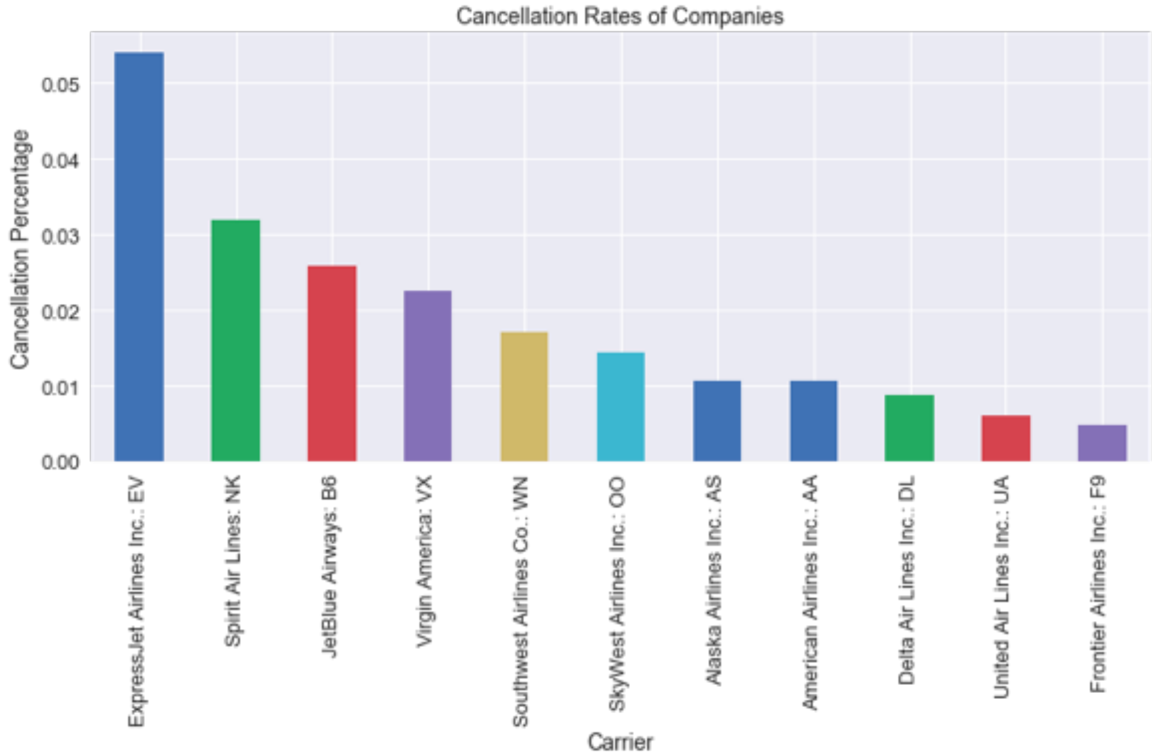
c. Flight Volumes by Companies:



The graphic shows the flight numbers for each Airline Company.

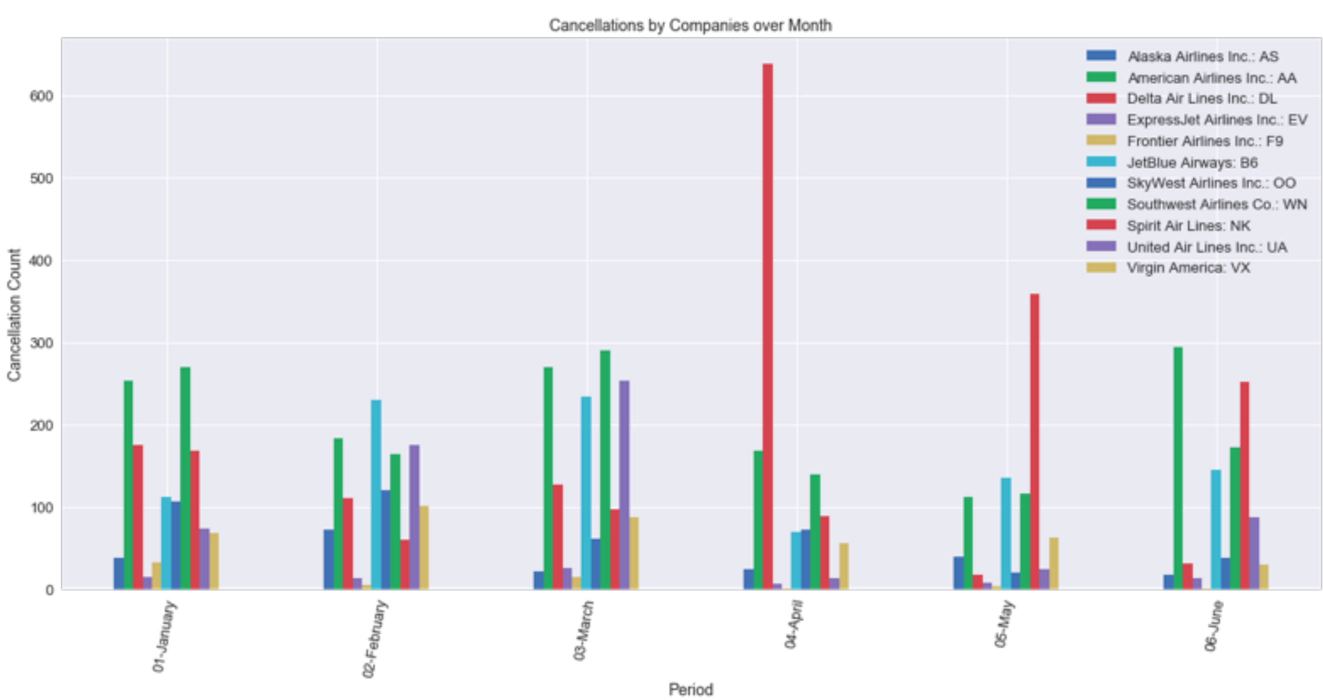
Delta Airline has the maximum flight operation while ExpressJest has the minimum.

d. Cancellations Rates by Companies:



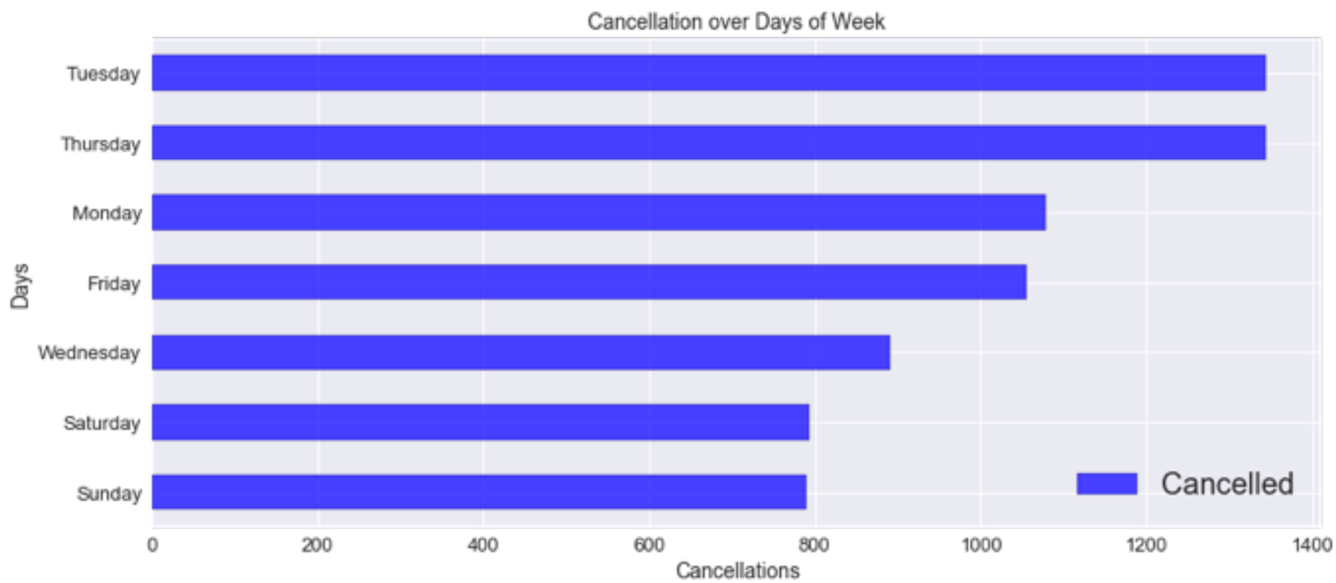
The graphic shows **cancellations rates** for each Airline Company. **Express Airline** has the maximum cancellation rate while **Frontier Airline** has the minimum. Big companies such as American Airlines, Delta Air or United Airlines' cancellation rates are less than 1%.

e. Cancellation by Companies over Months:



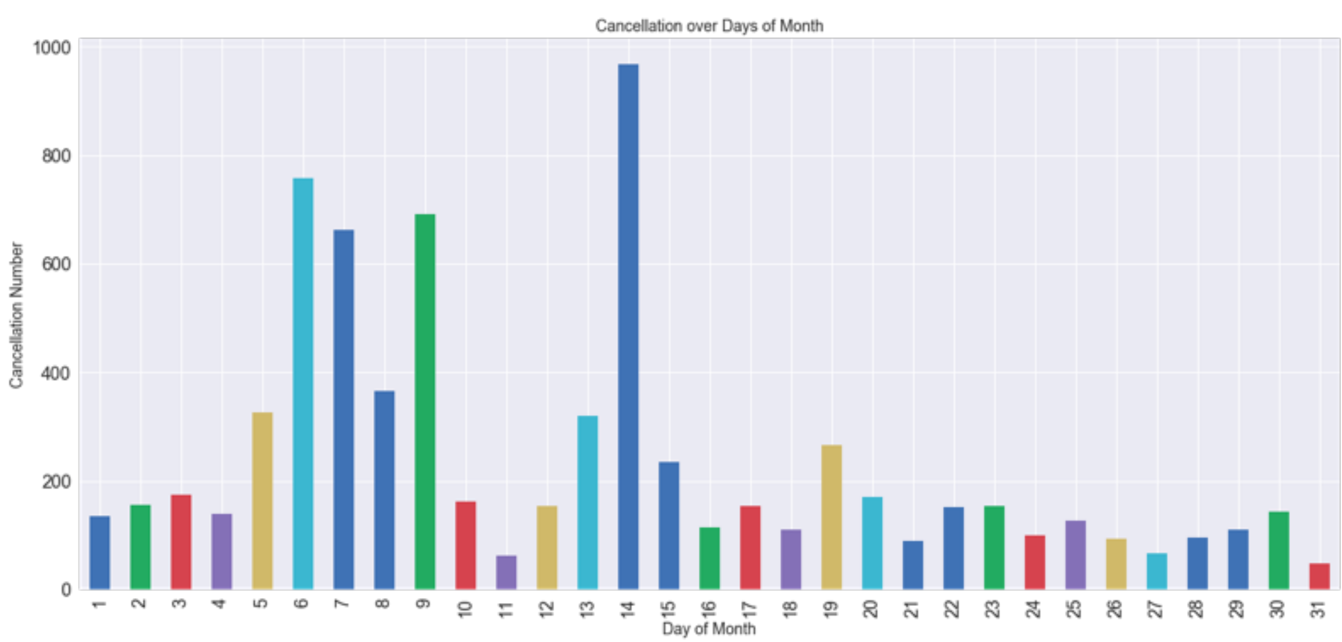
The graphic shows cancellation numbers of Airline companies over months. Most of the cancellations happened in March and April, especially Delta Airlines has many cancellations in April.

f. Cancellation Rates over Day of Week:



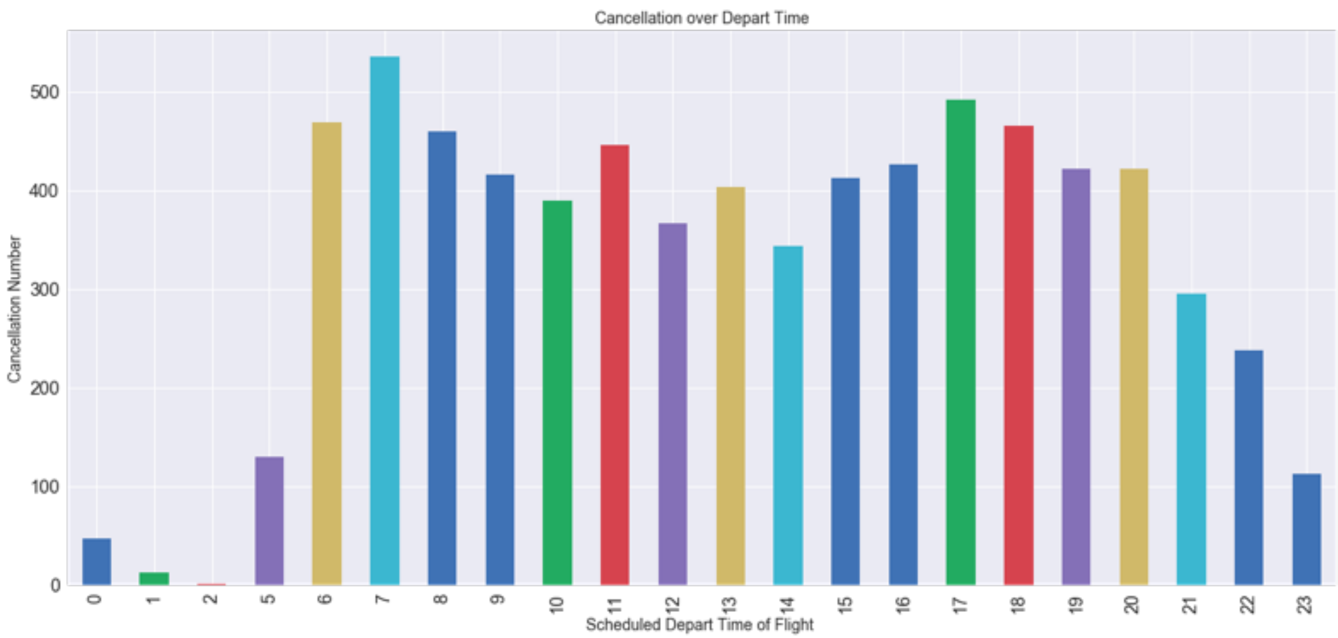
The graphic shows cancellation numbers for each day of week. As seen on the graphic, Weekdays have more cancellations than weekends. We added ***[PartOfWeek]*** new feature to our data set. We categorized the flights as weekday and weekend flights.

g. Cancellation Rates over Day of Month:



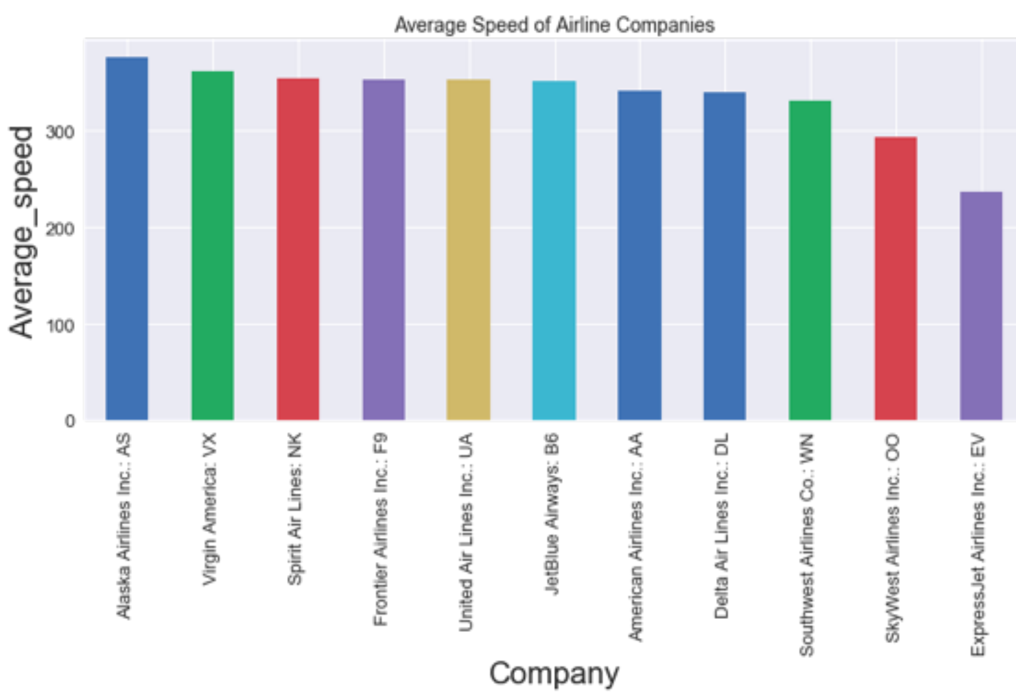
The graphic shows cancellation numbers for each day of month. First half of months have more cancellations than latter one. We added ***[PartOfMonth]*** new feature to our data set. We categorized the flights as first half and second half flights of month.

h. Cancellation Rates over Depart Hour:



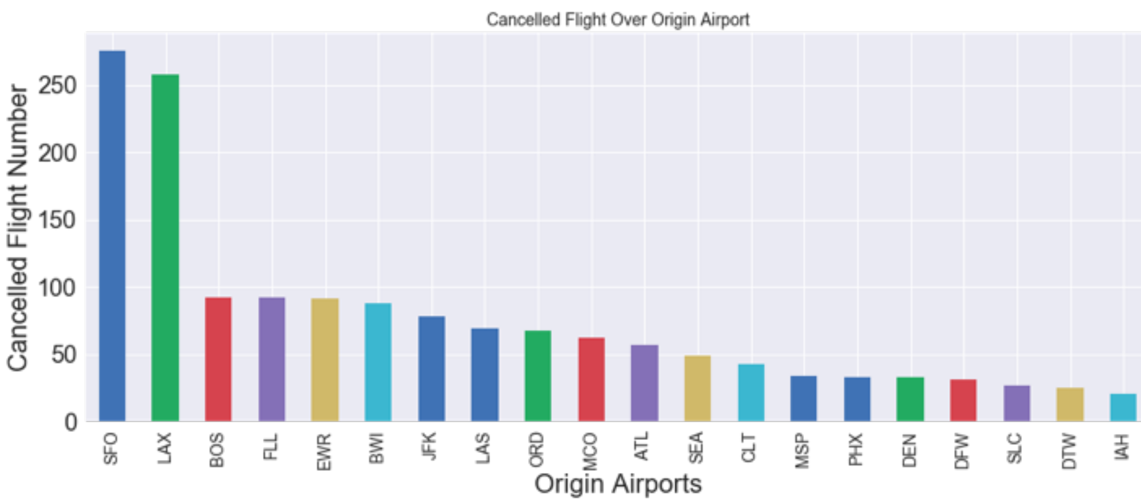
The graphic shows cancellation numbers for each hour. We noticed that most of the cancellations occurred between 6am -9 pm (21). We added ***[PartOfDay]*** new feature to our data set. We categorized the flights as 6am-9pm, 10pm-12 am and 1 am-5 am flights.

i. 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.

j. Cancelled Flights over Origin Airports



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

The graphic shows total cancelled flight numbers for each Origin Airport. **California** (Los Angeles and San Francisco) has the most cancellations

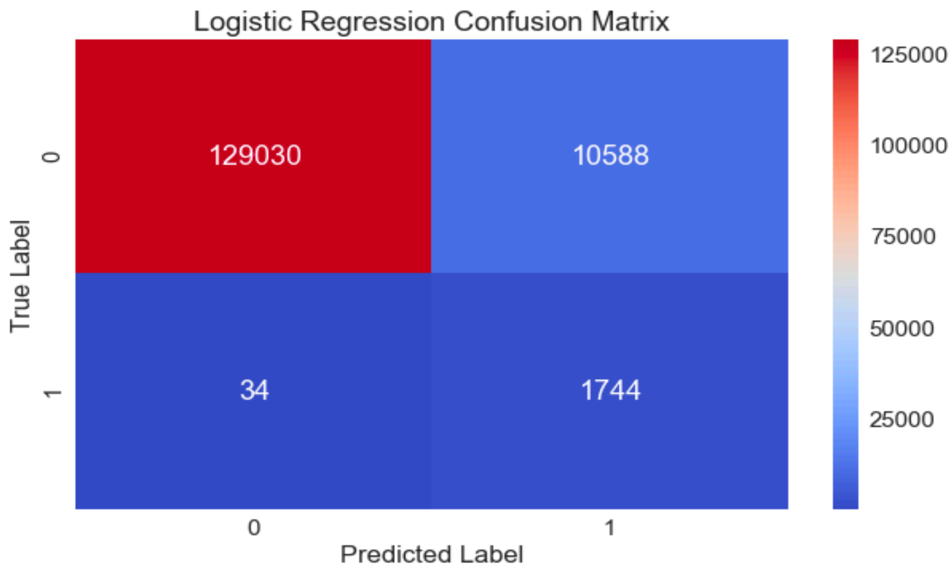
4. MACHINE LEARNING MODELS

This is a supervised binary classification problem. We are trying to predict the flights would be cancelled or not. We used Python's scikit learn libraries to solve our problem. In this context, we implemented Logistic Regression, k-Nearest Neighbors, Random Forest, Adaboost, and Gradient Boost algorithms.

We split our data set into training set (80%) and test set (20%). We converted our categorical data into numeric through label encoding and we used StandardScaler() to scale our data.

Additionally, we used 5 fold cross validation technique to get rid of overfitting problem. As a evaluation metric we used Area Under ROC Curve.

a. Logistic Regression:



Area Under Curve ROC: 0.9614

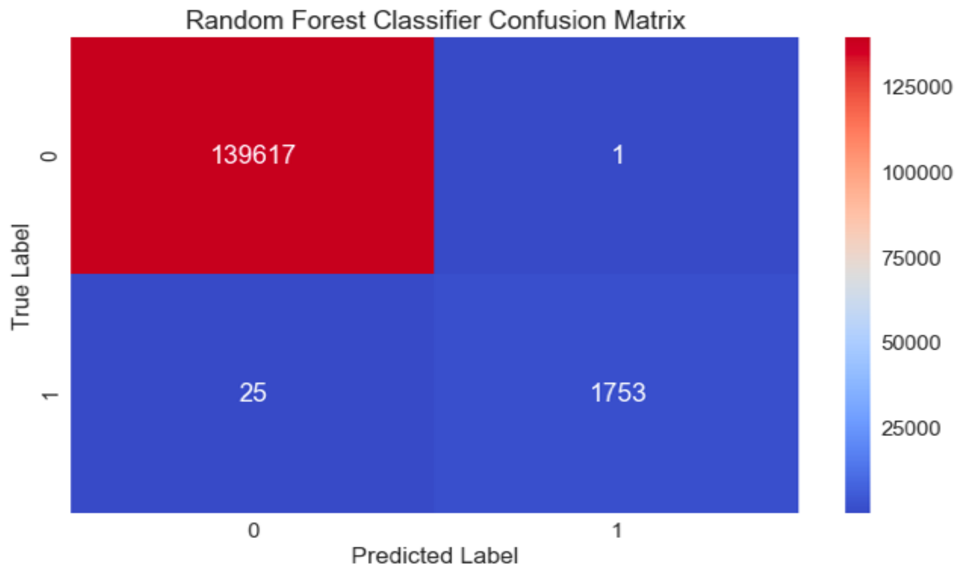
Logistic Regression is not working very well. It misclassified more than 10 thousand non-cancelled flights as cancelled. Namely, the False Positive Rate is incredibly high.

b. K-Nearest Neighbors:

Area Under Curve ROC: 0.9704

K-NN is working better than Logistic Regression. But it still misclassified almost 1/3 cancelled flights as non-cancelled, so Recall(sensitivity) rate (77%) is just moderate.

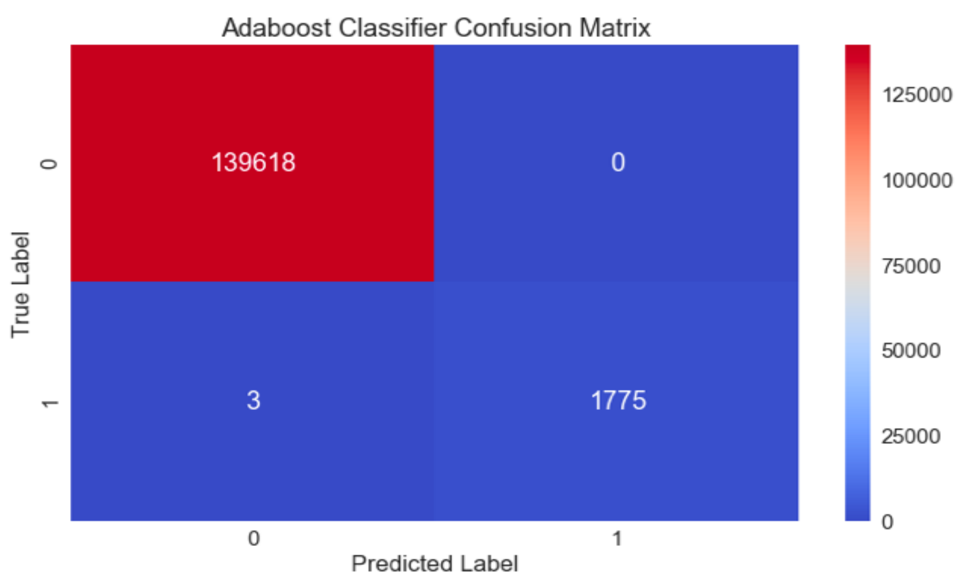
c. Random Forest:



Area Under Curve ROC: 0.9987

Random Forest Algorithm is working pretty well. better than Logistic Regression. It misclassified almost 1/3 cancelled flights classified as non-cancelled, so Recall(sensitivity) rate is just moderate.

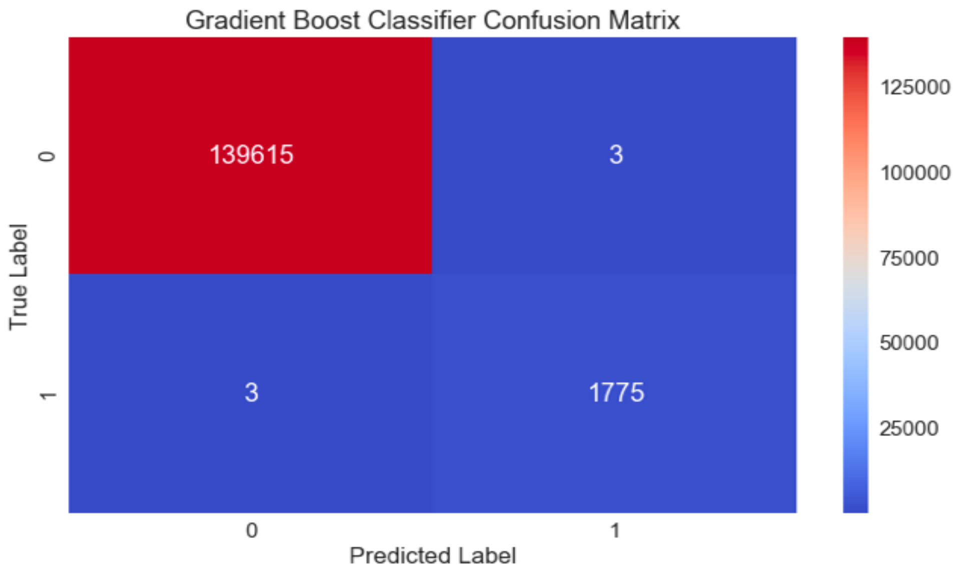
d. Adaboost:



Area Under Curve ROC: 0.999

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

e. Gradient Boost:



Area Under Curve ROC: 0.983

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.

f. Model comparison:

Model	AUC	Sensitivity (Recall)	Precision	Log Loss
Linear 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

Adaboost algorithm is the best algorithm while logistic regression algorithm is the worst one for our problem.

5. 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.

6. 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.