Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Prediction of Airline On-Time Performance

Executive Summary

The commercial airline industry is one of the most diverse and dynamic in the world. It is a capital intensive, labor intensive, highly regulated, and subject to the ever-changing customer demand. The market is extremely competitive with some extremely thin profit margins on large portions of airline networks. In order to remain competitive in the space, airlines must be able to attract and retain customers, and the fundamental component of gaining market share is having an on-time airline.

Arguably, the single most important metric observed by commercial airlines in known as on-time performance (OTP). OTP is a widely accepted method of understanding punctuality for different modes of public transport, not just aviation. It provides a standardized means of comparing how well one service provider operates according to its published schedule compared to another.

In the United States, the Department of Transportation defines a flight being on-time if they arrive within fourteen minutes of the original scheduled arrival time. This is colloquially known as the A14 rate (or A14 hit rate). Everything that an airline does focuses on maintaining an A14 rate determined acceptable by the executives of the company.

There are various methods of managing A14, but there must be a balance of performance metrics and profitability when designing and implementing upon a schedule. Additionally, airlines must build their networks to accommodate for seasonality, special events, and they must also be prepared to react to unexpected weather events, congestion, and other unexpected incidents. Not to mention increasing strain on airport passenger throughput, gating issues, and regulatory restrictions pose additional constraints on physical space within airports. However, the main issue lies in the fact that airlines want high confidence of on-time-performance throughout their network.

In the past, much of the network design was done by seasoned individuals with longstanding history of the company. Decisions would be made collectively, using a combination of history and individual professional opinions. These methods gradually grew more complex with the addition of the personal computer – using spreadsheets, programmatic algorithms, and event-modeling became extremely useful tools for these companies. Further, machine learning has become a boon for network planning.

Background / Context

Domain

The domain selected for the project is Airlines as described in the summary. This project aims to replicate the efforts of the commercial airlines by applying data mining and machine learning principles to determine important factors that drives airline OTP.

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Brief description of the scenario

Using historical data, can we determine what factors drive OTP? What are factors that lead to delayed or canceled airline flights? And, can these factors be used to predict on-time performance (OTP)?

The prime concern for us, as the project team, is the scenario where these airlines are not able to meet the OTP. Our focus would be the identification of factors affecting the OTP, and the opportunities available for improvement.

Decisions of interest

By using historical data, we plan to identify the biggest factors that influence on-time performance. Identify isolated areas that can be improved upon to optimize the OTP. Also, to discover the usefulness that can be gained from a classification/regression ML algorithm for factors affecting the OTP.

With these insights we hope to help the decision makers in their short term and long-term decisions related to the airline's performance without affecting the profitability.

Decision makers

Managers and associates in the network planning of airline companies would like to predict in advance if a flight will be on-time. This would help them in making primary decision on network planning and network design. They provide what, in their professional opinion, is the optimal, most recoverable network to fly. From there, it becomes the responsibility of the operations teams to implement upon the schedule. They must staff accordingly and plan for the operational readiness of each flight. Additionally, operations groups can benefit from knowing the likelihood of a late or cancelled flight. These would enable them to create fallback plans accordingly.

Also, this information can help executives to determine long-term strategy on where they would like to expand their airline, based on the networks that are most profitable and has the most likelihood for OTP.

Decision	Decision Maker	Details
Network Planning (Short Term)	Managers/Associates	Identify best performing networks with least affected OTP
Operations (Immediate Future)	Operations Managers/Groups	Be prepared for operational readiness and backup plans based on the insights
Executives (Long Term Strategic)	Senior Executives	Decisions on strategy as to how to best use the resources available in the operational networks; Publish schedules

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Business Understanding

Business Objective

The objective of this project is twofold. Firstly, use the data mining techniques learned to gain a fundamental understanding of what drives on-time-performance. Secondly, use the data to build a robust ML model to be able to predict both the likelihood of a flight being late (logistic regression classifier), and the time delta of being on-time or late (RF, NN, GB regression model). Using the information and tools provided by these models, a decision maker (e.g. operations manager) would theoretically be able to make proactive choices on how to assess real-time flight cancellations and re-accommodations.

We would like to address the following questions. What are the primary drivers of On-Time-Performance? What factors could an airline better manage to boost OTP?

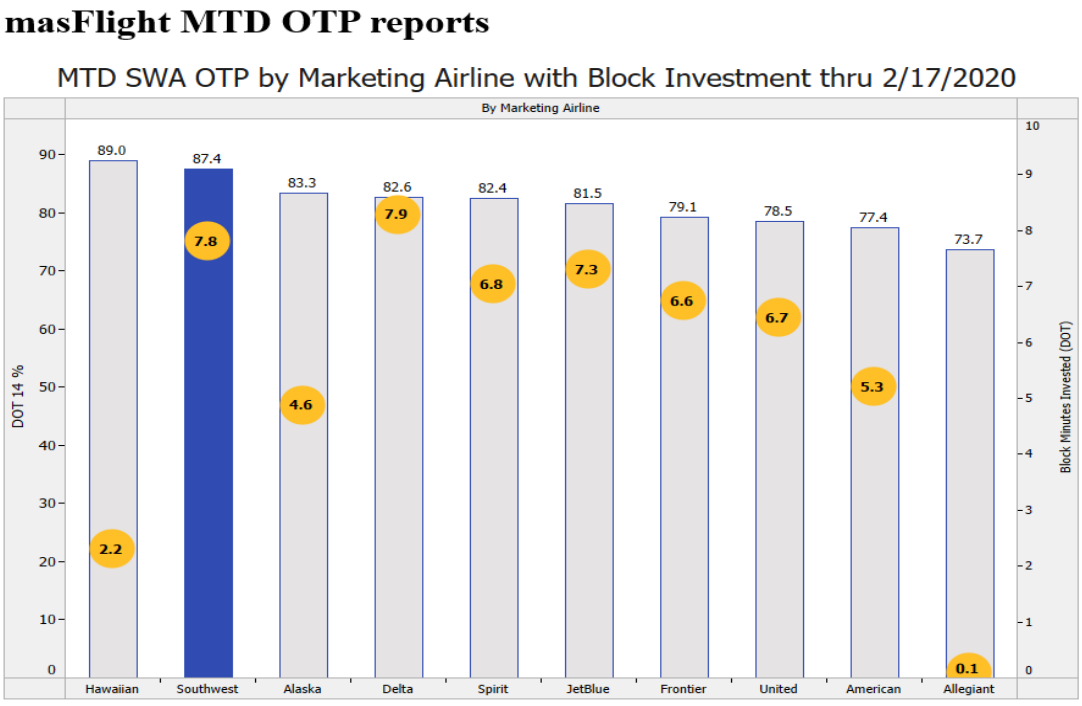


Figure 1: A month-to-date view of On-Time-Performance of various airlines for a day. These daily metrics are aggregated for reporting for responsible managers and executives to assess airline operation performance.

Situation Assessment

Executives and network decision-makers of airline companies would like to know in advance if a particular flight will be on-time. Using historical data, can we determine what factors drive OTP? What

are factors that lead to delayed or canceled airline flights? And, can these factors be used to predict on-time performance (OTP)? Answers to these questions will be invaluable for real-time decision making in the aviation world. With reliable predictions of OTP for a given flight, managers and directors can decide whether to keep the flight and take a hit to overall customer satisfaction on OTP, or rather to cancel the flight and reaccomodate the passengers and take the hit to OTP. These are questions that operations teams face every day.

Data Mining Goals

The goals of this project are to determine relationships (if any) between weather information, historical performance, and real on-time-performance. With these relationships in mind, can we build a tool that could potentially provide real-time information on the chance of a delayed flight and the severity of delay.

Data Understanding

Data requirements

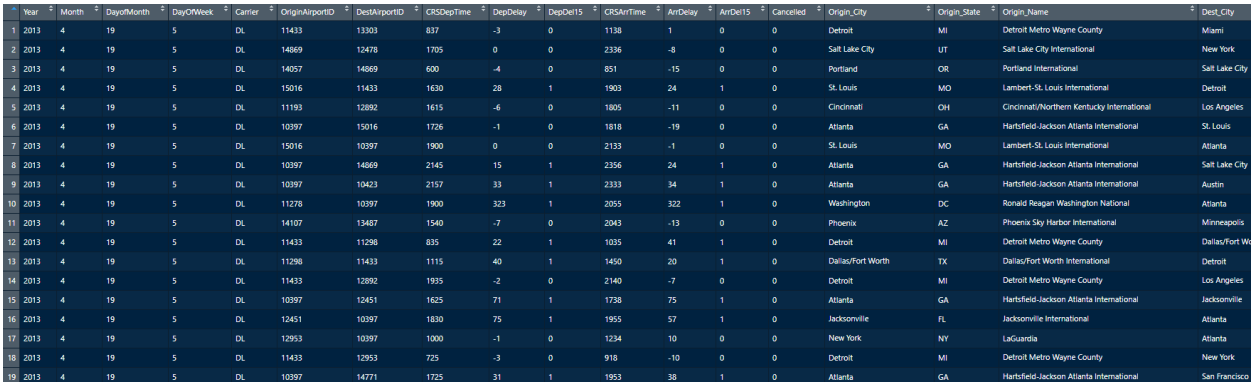
For this project to be meaningful, we must have a large dataset going back several years to effectively apply data mining techniques to derive the insights. The data must be enough for a machine learning model of reasonable complexity (e.g. a neural network) to arrive at excellent predictive results of accuracy, precision, and recall for classification, and reasonable MAU, RMSE, and MAPE for regression. With over 87000 flights per day in the USA, which are regularly tracked, this should not be a problem.

Data required to proceed with the project are the flight records and their on-time performance details. Additionally, weather details for different stations from the past (historical data) and details regarding airports are needed as well. There are 3 data tables that are being considered for the project. First one deals with recorded flights in the past and the associated delays. The second table consists of the list of the airport details. Third one lists the weather condition at the departing and arriving airports at points of time.

Describe data

The data from the DOT on commercial airlines includes historical flight data. The fields include year, month, day, carrier, origin airport, destination, airport, scheduled departure and arrival times, actual departure and arrival times. The specific data tables being used, and the fields along with their attributes are given in detail below. Some aggregate views of recent data can be seen throughout this proposal.

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez



	Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirportID	DestAirportID	CRSDepTime	DepDelay	DepDel15	CRSArrTime	ArrDelay	ArrDel15	Cancelled	Origin_City	Origin_State	Origin_Name	Dest_City
1	2013	4	19	5	DL	11433	13303	837	-3	0	1138	1	0	0	Detroit	MI	Detroit Metro Wayne County	Miami
2	2013	4	19	5	DL	14869	12478	1705	0	0	2336	-8	0	0	Salt Lake City	UT	Salt Lake City International	New York
3	2013	4	19	5	DL	14057	14869	600	-4	0	851	-15	0	0	Portland	OR	Portland International	Salt Lake City
4	2013	4	19	5	DL	15016	11433	1630	28	1	1903	24	1	0	St. Louis	MO	Lambert-St. Louis International	Detroit
5	2013	4	19	5	DL	11193	12892	1615	-6	0	1805	-11	0	0	Cincinnati	OH	Cincinnati/Northern Kentucky International	Los Angeles
6	2013	4	19	5	DL	10397	15016	1726	-1	0	1818	-19	0	0	Atlanta	GA	Hartsfield-Jackson Atlanta International	St. Louis
7	2013	4	19	5	DL	15016	10397	1900	0	0	2133	-1	0	0	St. Louis	MO	Lambert-St. Louis International	Atlanta
8	2013	4	19	5	DL	10397	14869	2145	15	1	2356	24	1	0	Atlanta	GA	Hartsfield-Jackson Atlanta International	Salt Lake City
9	2013	4	19	5	DL	10397	10423	2157	33	1	2333	34	1	0	Atlanta	GA	Hartsfield-Jackson Atlanta International	Austin
10	2013	4	19	5	DL	11278	10397	1900	323	1	2055	322	1	0	Washington	DC	Ronald Reagan Washington National	Atlanta
11	2013	4	19	5	DL	14107	13487	1540	-7	0	2043	-13	0	0	Phoenix	AZ	Phoenix Sky Harbor International	Minneapolis
12	2013	4	19	5	DL	11433	11298	835	22	1	1035	41	1	0	Detroit	MI	Detroit Metro Wayne County	Dallas/Fort W
13	2013	4	19	5	DL	11298	11433	1115	40	1	1450	20	1	0	Dallas/Fort Worth	TX	Dallas/Fort Worth International	Detroit
14	2013	4	19	5	DL	11433	12892	1935	-2	0	2140	-7	0	0	Detroit	MI	Detroit Metro Wayne County	Los Angeles
15	2013	4	19	5	DL	10397	12451	1625	71	1	1738	75	1	0	Atlanta	GA	Hartsfield-Jackson Atlanta International	Jacksonville
16	2013	4	19	5	DL	12451	10397	1830	75	1	1955	57	1	0	Jacksonville	FL	Jacksonville International	Atlanta
17	2013	4	19	5	DL	12953	10397	1000	-1	0	1234	10	0	0	New York	NY	LaGuardia	Atlanta
18	2013	4	19	5	DL	11433	12953	725	-3	0	918	-10	0	0	Detroit	MI	Detroit Metro Wayne County	New York
19	2013	4	19	5	DL	10397	14771	1725	31	1	1953	38	1	0	Atlanta	GA	Hartsfield-Jackson Atlanta International	San Francisco

Figure 2: Dataset for Flight Delays

Data tables and their attributes along with their relevance in the scope of this data mining project are discussed below.

Sl. No:	DataTable: Flight_Delays	Data Mining Relevance
1	Year (Integer)	Date
2	Month (Integer)	Date
3	DayofMonth (Integer)	Date
4	DayOfWeek (Integer)	Date
5	Carrier (String)	Company Name/Identity
6	OriginAirportID (String)	Geographic/Identity Information
7	DestAirportID (String)	Geographic/Identity Information
8	CRSDepTime (timestamp)	Date-time
9	DepDelay (minutes)	Predictor information
10	DepDel15 (Boolean)	Predictor information
11	CRSArrTime (timestamp)	Scheduled arrival time
12	ArrDelay (minutes)	Objective (y-test/y-predict)
13	ArrDel15 (Boolean)	Objective (y-test/y-predict)
14	Cancelled (Boolean)	Predictor information
15	Origin_City (String)	Geographic/Identity Information
16	Origin_State (String)	Geographic/Identity Information
17	Origin_Name (String)	Geographic/Identity Information
18	Dest_City (String)	Geographic/Identity Information
19	Dest_State (String)	Geographic/Identity Information
20	Dest_Name (String)	Geographic/Identity Information

Sl. No:	DataTable: Weather	Data Mining Relevance
1	City (String)	Geographical Information
2	Temp (C)	Weather Information
3	Dew Point (C)	Weather Information
4	Wind Speed (knots)	Weather Information
5	Wind Gust (knots)	Weather Information
6	Wind Direction (360)	Weather Information

7	Weather_Desc (String)	Weather Information
8	Cloud Conditions (String)	Weather Information
9	Visibility (String)	Weather Information
10	Sky conditions (String)	Weather Information

Sl. No:	Airport Codes	Data Mining Relevance
1	Airport_id (String)	Identity Information
2	City (String)	Geographic/Identity Information
3	State (String)	Geographic/Identity Information
4	Name (String)	Identity Information

Sources

The data that is being used for this project is sourced from publicly available data sources. The Department of Transportation keeps a variety of data stores for all airlines that operate within the United States, and there are numerous aviation regulatory authorities across the globe that track commercial flight operations within their respective countries. Our data is sourced from masFlight – a SaaS aviation software company that specializes in the collection and analysis of large amounts of commercial aircraft operational data globally, from sources such as global flight information systems, schedules, ADS-B and proprietary information sources. The weather data is collected from the National Oceanic and Atmospheric Administration (NOAA) and the National Weather Service (NWS). As the data is provided by a 3rd party and lies behind a paywall, the data was acquired via an airline company source.

The links and resources to the datasets being used in the project are listed below for easy access and reference. Global Eagle is a paid service that provides the aggregated information, and the below sources would provide raw data inputs.

1. http://masflightbigdata.com/index.php & https://www.globaleagle.com/
2. https://www.aviationweather.gov/metar
3. https://www.aviationweather.gov/

Quality

The data is being sourced through a US government organization, The Department of Transportation, and a third-party aggregator (masFlight). Private sources are grounded in DOT data, and regularly scrubbed for accuracy and cleanliness. The weather data that is being sourced through a private data repository (masFlight) which provides high quality weather data for major airlines in the US. This company regularly screens their data and ensures the credibility and quality of data regarding the weather. Many major US carriers utilize these resources for analyzing flight performance. With regards to the airport details, these are available from multiple publicly data sources such as Kaggle and the data meets high quality standards. A snapshot of the weather data can be seen in the figure below.

origint emp	originde wpoint	originwind direction	originwin dspeed	originwi ndgust	originvi sibility	originwxstrin g	originskyc ondition1	originclou dlevel1	originskyc ondition2	originclou dlevel2	originskyc ondition3	originclou dlevel3	destte mp	destde wpoint	destwind direction	destwin dspeed	destwin dgust	destvis ibility	destwxstrin g	deststkyco ndition1	destclou dlevel1	deststkyco ndition2	destclou dlevel2	deststkyco ndition3	destclou dlevel3
12.20	5.60	20	9	0	10	-	BKN	25000	-	-	-	-	21.70	9.40	70	15	21	10	-	-	-	-	-	-	-
27.20	20.60	220	8	0	10	-	BKN	3000	BKN	25000	-	-	12.80	4.40	250	13	-	10	-	SCT	4300	BKN	20000	-	-
9.40	5.00	0	0	0	8	-	FEW	25000	-	-	-	-	12.20	7.80	320	17	24	10	-	FEW	1500	BKN	6000	OVC	8000
18.90	5.00	90	5	0	10	-	BKN	25000	-	-	-	-	21.10	20.60	220	8	-	2	+TSRA BR	FEW	800	BKN	2200	OVC	3200
23.90	17.80	240	8	0	10	-	BKN	25000	-	-	-	-	28.90	23.30	90	16	23	10	-RA	SCT	2100	SCT	5000	-	-
30.60	22.20	170	5	0	10	-	FEW	4500	FEW	8000	SCT	13000	27.20	23.30	80	10	-	10	-	SCT	2300	SCT	7000	-	-
25.00	23.30	0	0	0	10	-	FEW	6500	BKN	18000	-	-	23.30	22.80	0	0	-	10	-	FEW	5000	SCT	7000	-	-
26.70	20.60	60	11	0	10	-	FEW	2000	FEW	5000	SCT	8000	25.00	21.70	60	12	-	10	-	FEW	1900	BKN	3600	OVC	4700
-2.20	-5.60	0	0	0	10	-	-	-	-	-	-	-	-0.60	-3.30	0	0	-	10	-	FEW	25000	-	-	-	-
11.00	7.00	350	11	0	15	-	FEW	1900	BKN	10000	OVC	12000	1.00	-3.00	190	4	-	40	-	SCT	14000	BKN	20000	-	-

Figure 5: Snapshot of weather data from NOAA/NWS via masFlight.

Additionally, we ensure that our dataset contains relevant information for all predictor columns. (E.g. flight date is between 2013 and 2020, timestamps for take-off and landing are valid, block time and scheduled block time are not null, and time on-gate are not null.)

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Data Preparation

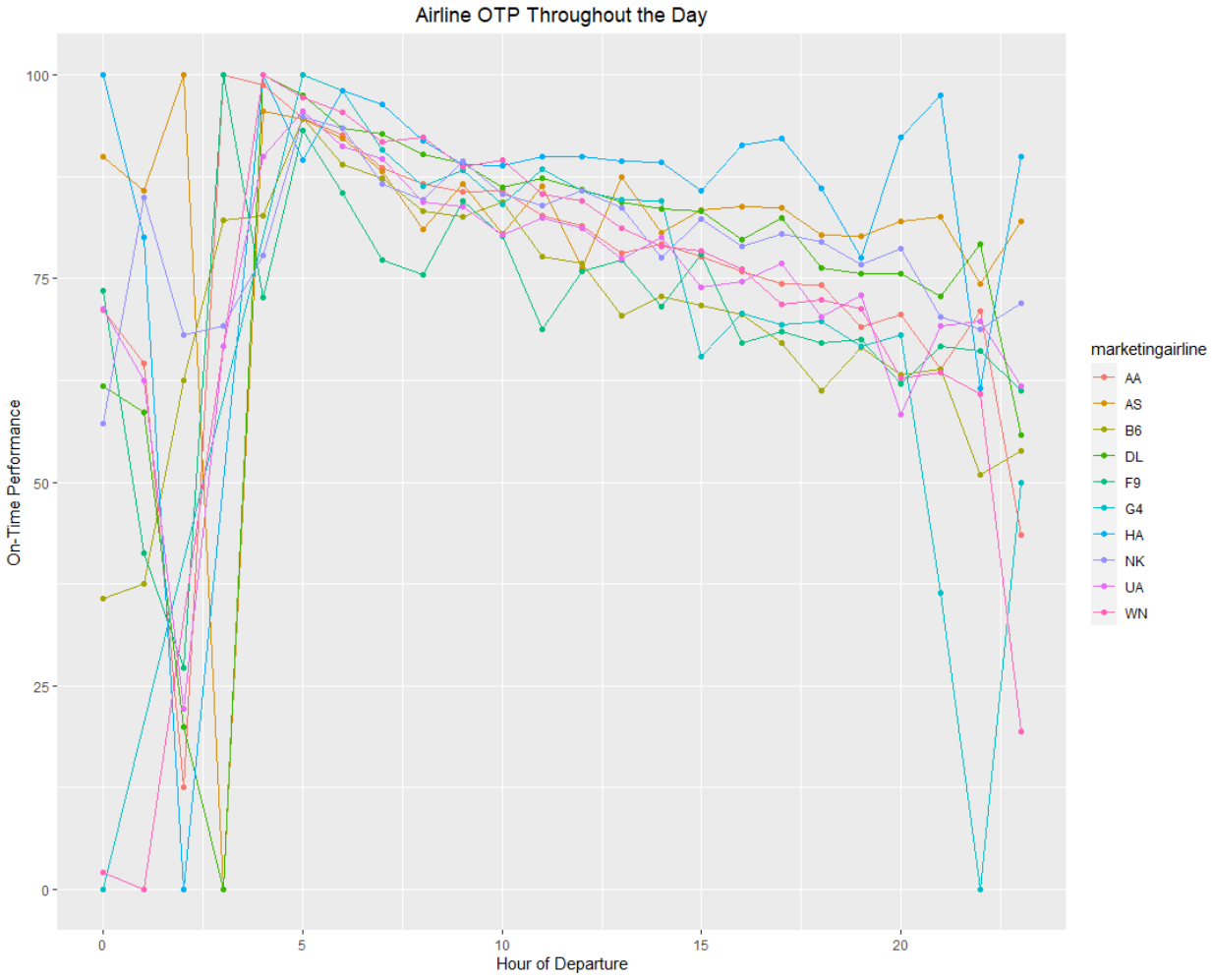
Data selection

Identifying attributes that had data of minimum value to the model and discarding them was a key part of this phase. Adding non relevant attributes would affect the quality of the model that is being build. The primary goal was to identify the critical columns from the data source that are of high significance to the model.

We used various visualization plots as part of the data exploratory phase to identify the relevance and relationship between the data attributes. The plots and their basic explanations are given below.

How OTP (% of on-time flights) looks like throughout the day (Grouped by marketing airline)

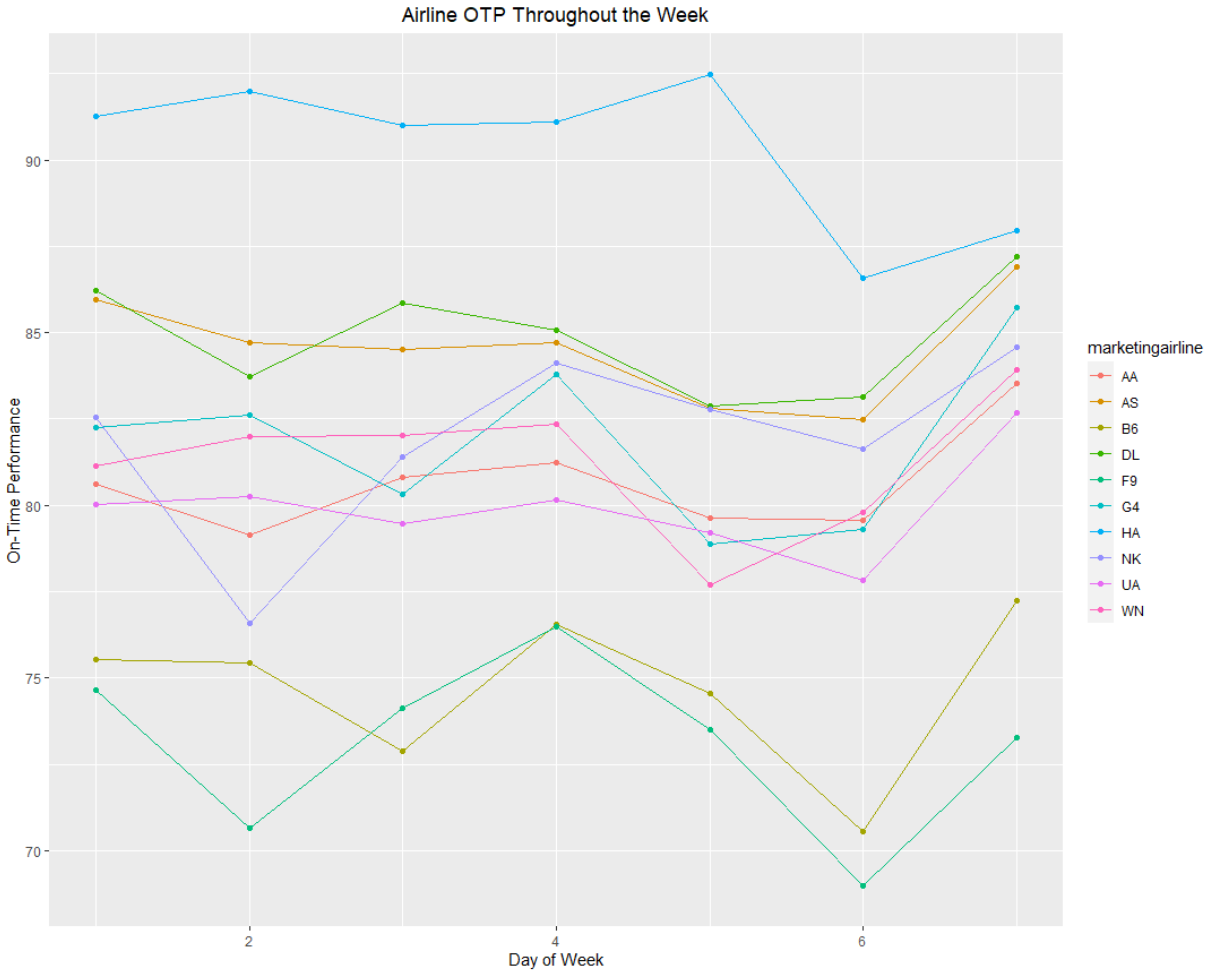
You could clearly see how most of the airlines have a high OTP around 5AM, and comparatively bad OTP performance around 2AM and 3AM.



Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

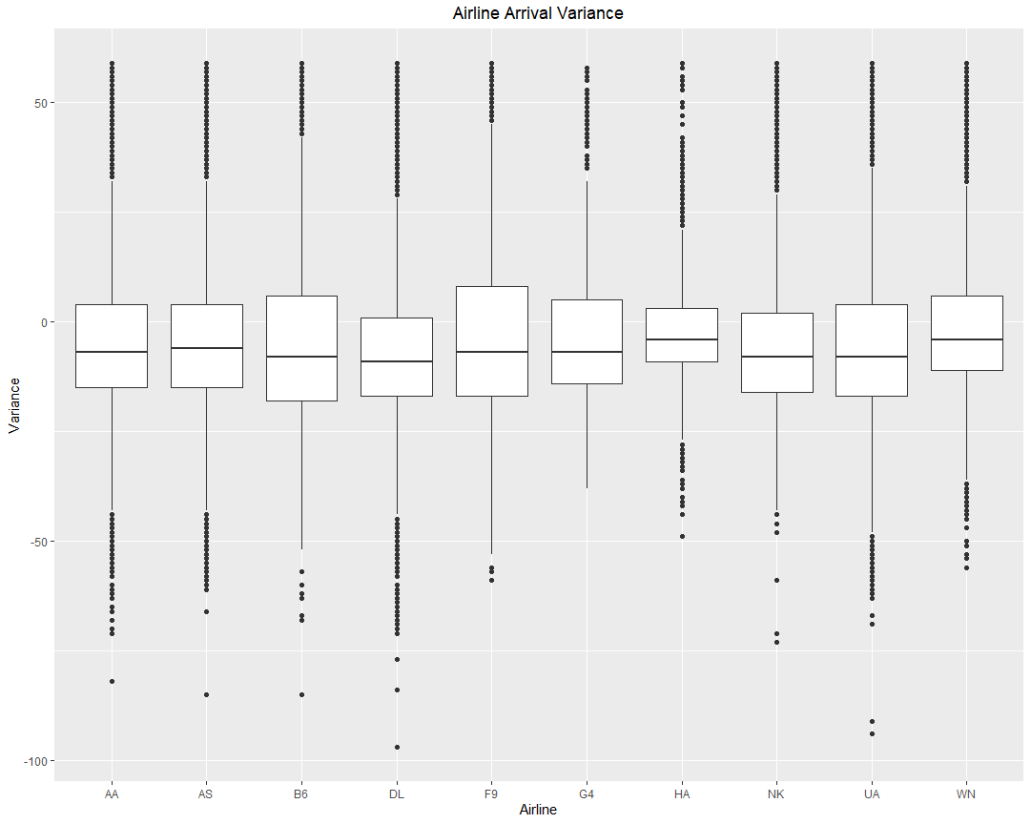
What does OTP (% of on-time flights) look like by day of the week (Grouped by marketing airline)

Generally, the airlines have their best OTP during Day1 of the week. Most of the airlines have their OTP getting dropped towards the Day6 of the week followed by a sharp rise in OTP towards Day7.

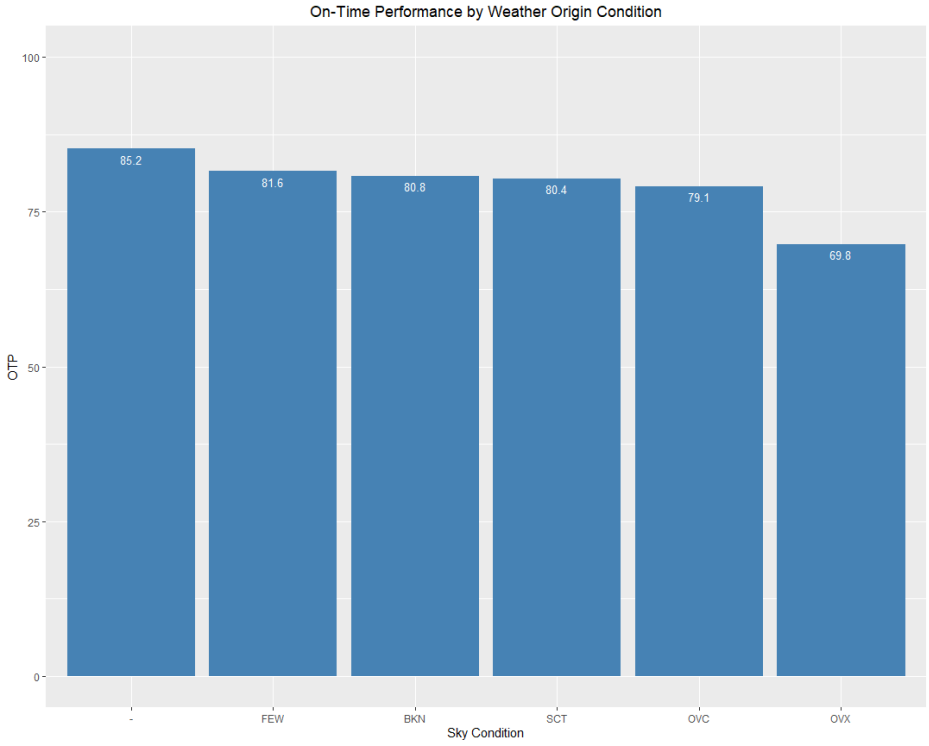


Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Boxplots of Airline Arrival Variance (Actual - Scheduled) Grouped by airline

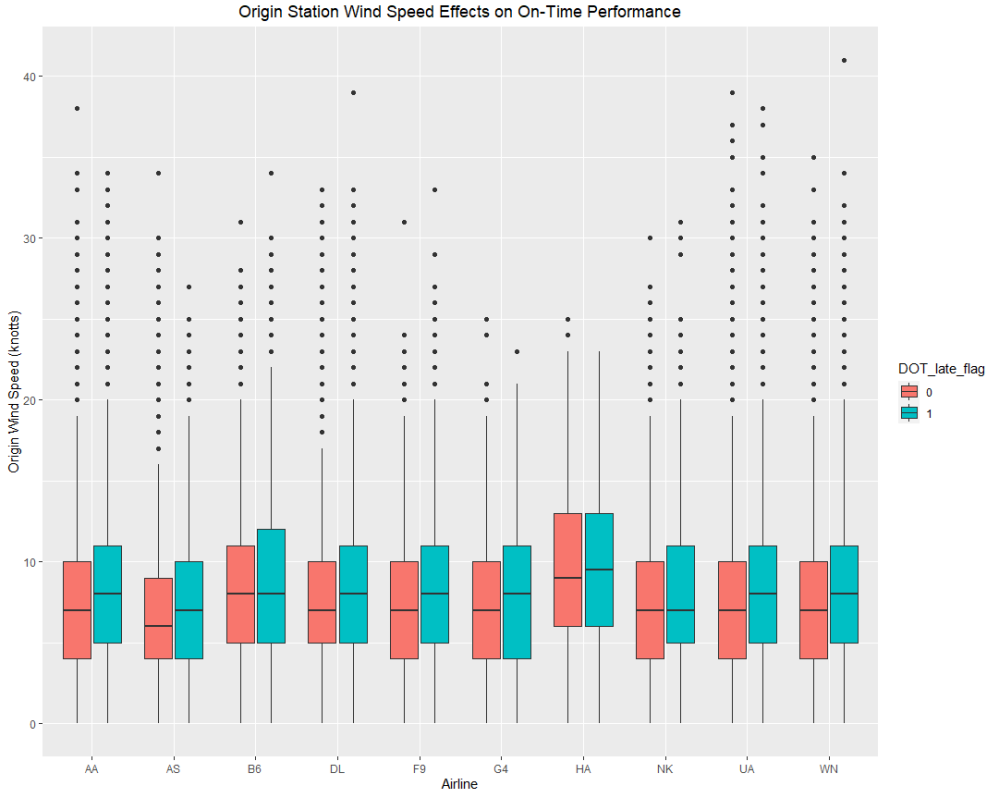


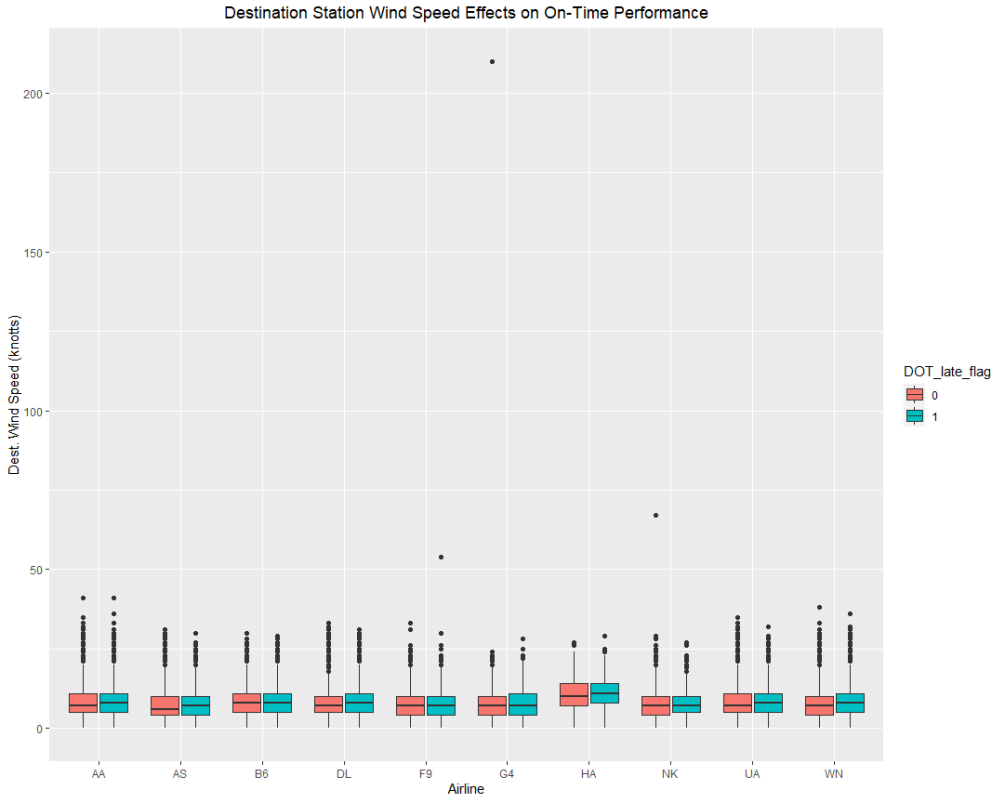
OTP Grouped by Weather Category



Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

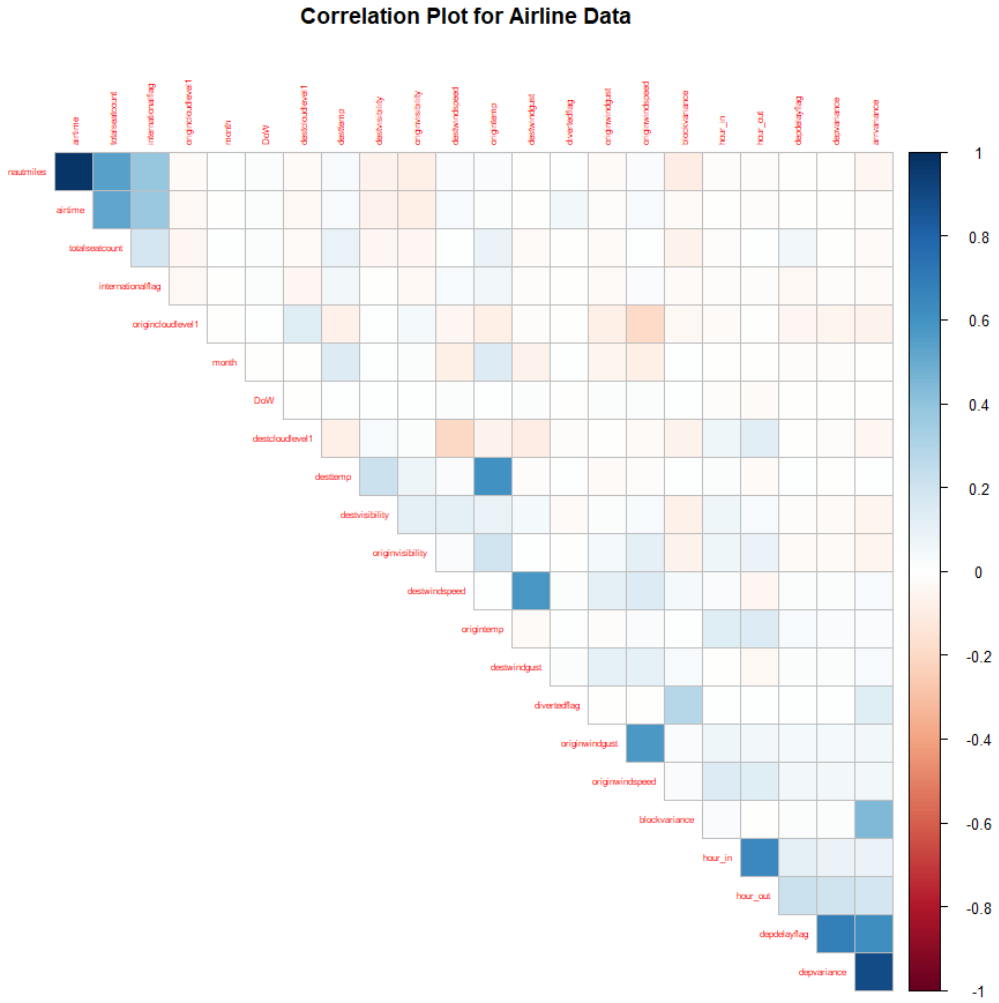
Effect of Wind Speed on OTP





Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Correlation Plot



The data exploratory phase visualizations helped us identify the columns that were least critical to the model. The removed attributes are 'arrvariance', 'late_flag', 'blockvariance' which were duplicate attributes of the data already available in the data set based on these visualizations. The initial data selection (from the original dataset) was completed based on the discussions with subject matter experts. The recommendations and suggestions from subject matter experts played a crucial role in the data selection phase.

Data cleaning

Data cleaning, being one of the critical phases that decides the success of the project was a priority for the team. We ensured that the data being pulled from the sources had minimum NULL values so that the integrity of the data is not affected. This was specifically enforced by pulling data from databases with explicit conditions to not include NULL values for critical data fields. Also, we had decided to use the OTP performance details of 9 airlines specifically that has the minimum missing data to ensure

minimum lead times on data cleaning activities. Also, we had ensured that the data has enough records to efficiently support a classifier model.

Prepare Data

In order to prepare data for the model building, we identified and removed columns that were having similar/duplicating properties of other columns in the data set. Also, numeric fields that had missing values were imputed with the value as 0. This was done based on the observation that the time variance was normally distributed (50% of the flights reach before time and 50% flights are delayed). Zero would be the mean/median in a normally distributed scenario with equal positive and negative variations.

Some of the data fields were converted to factors and some other fields were converted to numeric fields based on their relevance and nature of the data. Also, we had used one-hot encoding to convert categorical data into dummy values.

Also, fields that were not relevant were removed from the data frame before model building. The dataset was divided into training data (60%) and validation data (40%).

Modeling

Describe data in detail

The cleaned and prepared data was divided into two data sets. The training dataset had 120000 rows (60%) and the validation dataset contained 80000 rows (40%). Dataset had enough records to efficiently support a classification model. Much of the data was cleaned in the data pull (SQL query), and had minimal noise, and was ready for the decision of choice of model for this classification problem.

The features of interest were the month, day of week, the marketing airline, origin, destination, seat count, distance, time of departure, time of arrival, diverted flag, windspeed, temperature, visibility, and time-on-gate variance. Other fields used were transformations of these columns (e.g. dummy variables). The target variable was a DOT late flag (i.e. arriving 15 minutes or more after the scheduled arrival time).

What type of decision-making models are appropriate for the decision-making tasks?

The decision-making model suitable for the task at hand, which is to predict if a flight would be on-time or delayed was a classification model. A classification model will be able to predict whether a flight will meet its on-time performance or not. The team used ‘Logarithmic Regression Classification’ model and a ‘Random Forest Classification’ as part of the project. These models were selected for their simplicity and ease of explanation for higher management.

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Provide rationale for choice of models

We decided to use these two models to have a basic comparison between the predictions. The relationship between the variables did not look like they were linearly related which led us to take the logarithmic regression approach in the first place. The random forest method was decided to support and provide a comparison for the model already selected. The logarithmic regression model which is quite complex, being used along with a random forest model with more interpretability and transparency was the most logical and practical approach. These approaches are valued for their simplicity and ease of explanation for higher management.
We could have used a Neural network model for better predictions. However, the enormous data that ruled out the possibility to implement and try a neural network model with the hardware and resources that we had. Training time for a neural network exceeded 48 hrs.

Detail model development and output

Logistic Classification

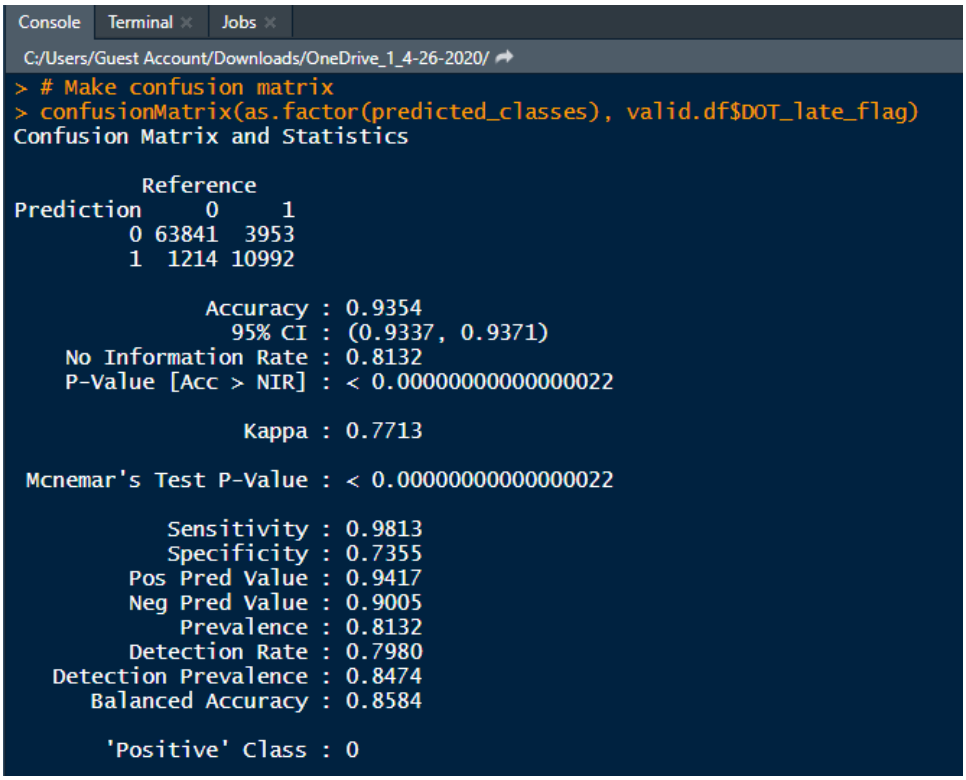
The data were prepared to have continuous fields for the numeric columns, and dummy columns (i.e. one-hot encoding) were done for the nominal categorical variables such as origin and destination. These fields were then passed into the logistic regression algorithm for model training. The GLM package was used.

Accuracy Summary

```
> # Summarize accuracy
> table(predicted_classes, valid.df$DOT_late_flag)

predicted_classes     0     1
                0 63841  3953
                1  1214 10992
>
```

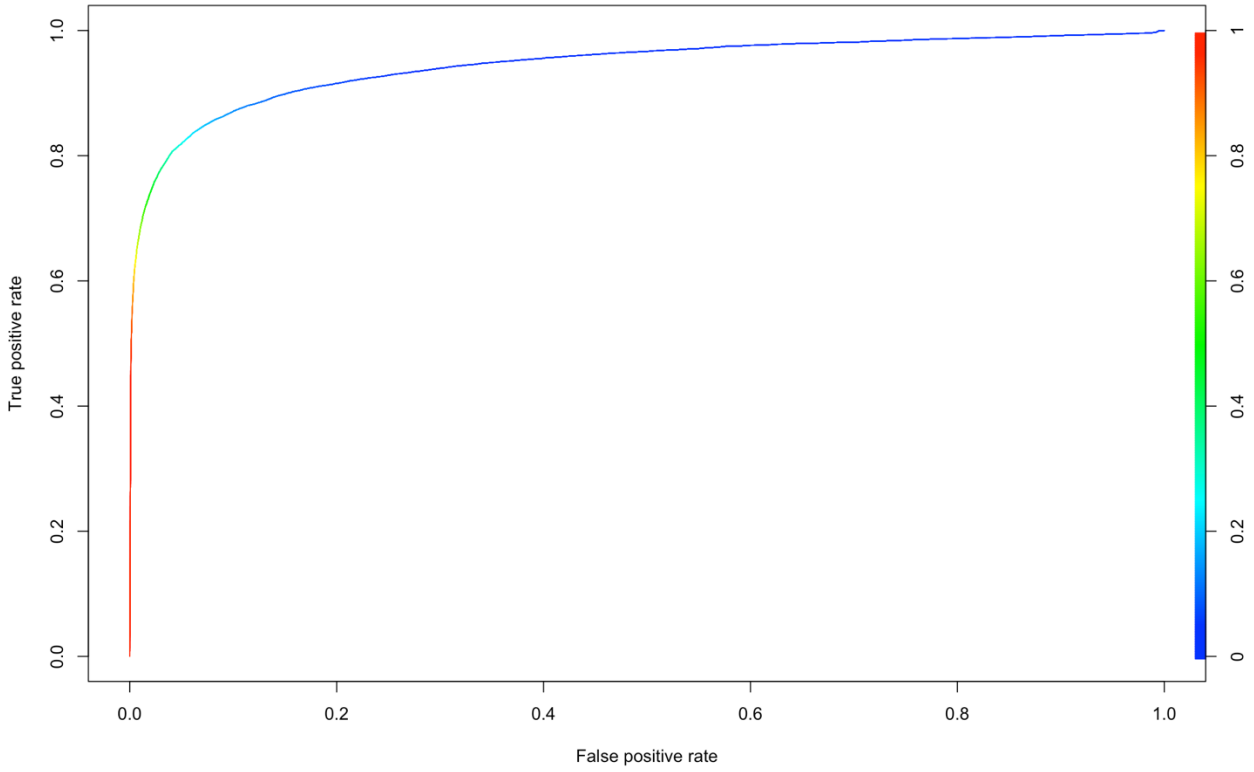
Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Confusion Matrix



Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

ROC Curve – AUC = 0.944

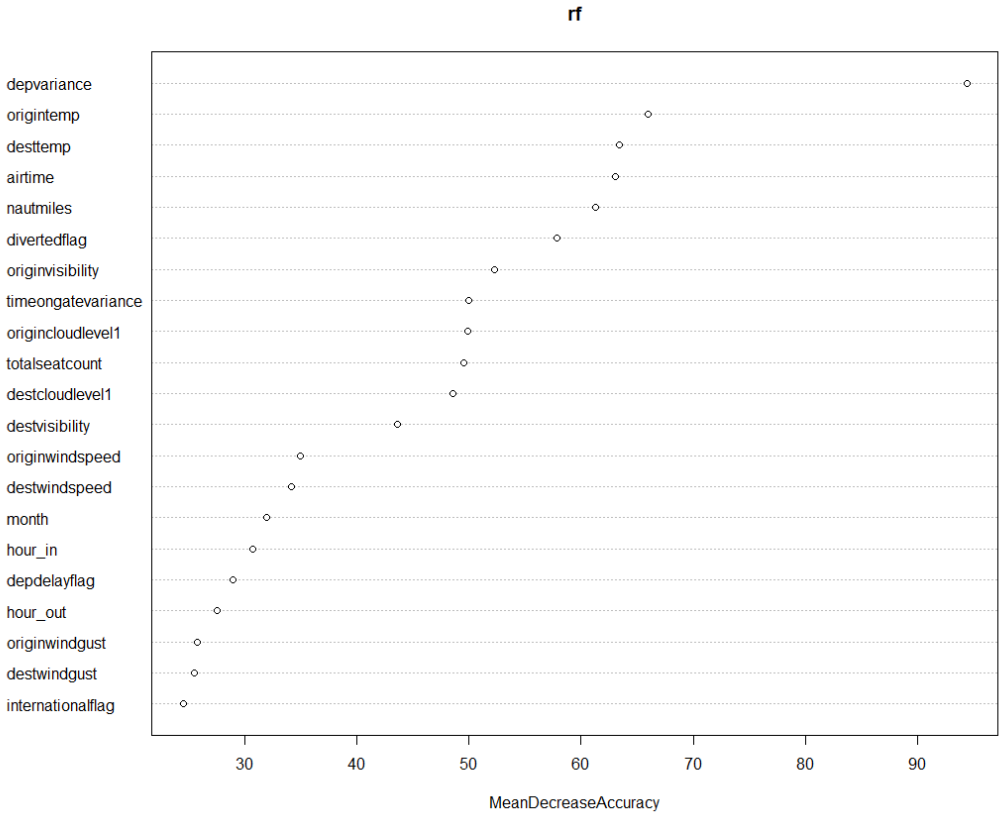


Random Forest Classifier

Similarly to logistic regression, the data were prepared to have continuous fields for the numeric columns, and dummy columns (i.e. one-hot encoding) were done for the nominal categorical variables such as origin and destination. These fields were then passed into the random forest classifier algorithm for model training. The randomForest library was used. The max depth parameter was set to 2, and all else was left as the default.

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Variable Importance Plot



Accuracy Summary

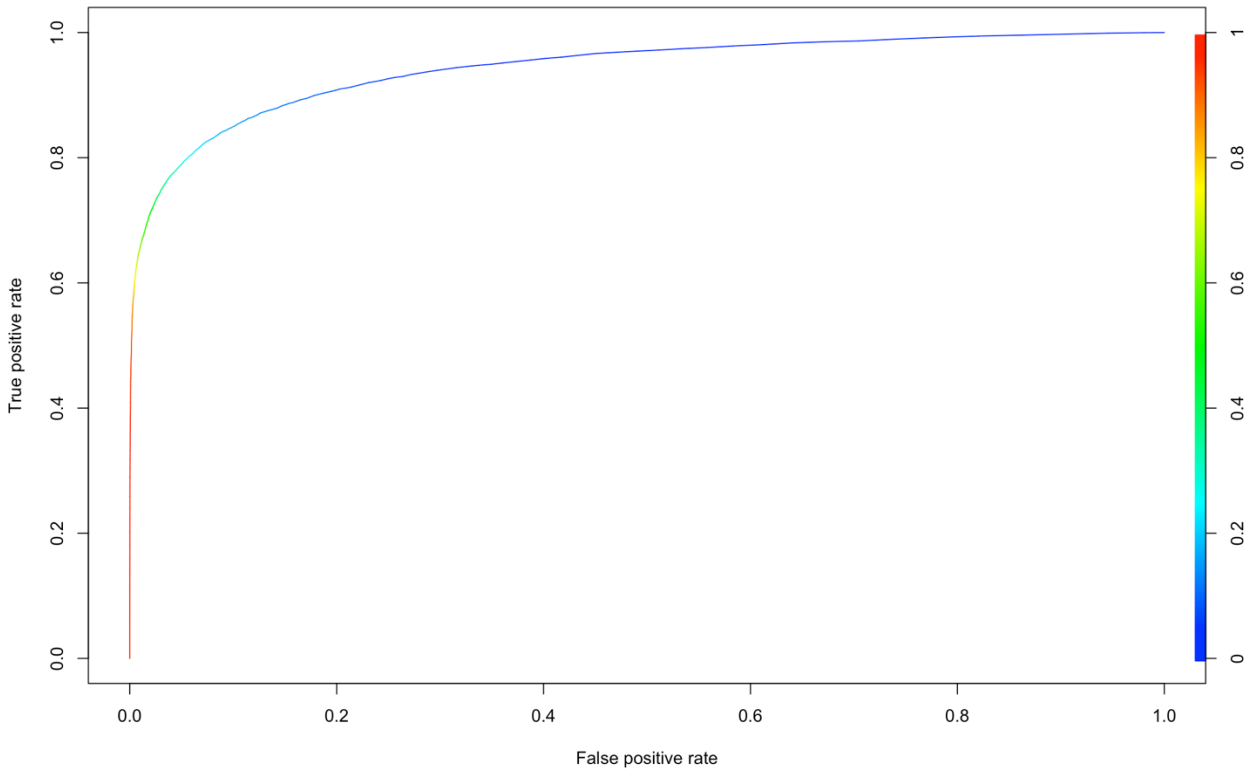
```
> # Make predictions
> rf.pred <- predict(rf, valid.df)
> # Summarize accuracy
> table(rf.pred, valid.df$DOT_late_flag)

rf.pred     0     1
      0 63823  4369
      1  1232 10576
```

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Confusion Matrix

```
> # Make confusion matrix
> confusionMatrix(rf.pred, as.factor(valid.df$DOT_late_flag))
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 63823  4368
         1  1232 10577

               Accuracy : 0.93
                 95% CI : (0.9282, 0.9318)
    No Information Rate : 0.8132
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7493

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9811
            Specificity : 0.7077
         Pos Pred Value : 0.9359
         Neg Pred Value : 0.8957
             Prevalence : 0.8132
         Detection Rate : 0.7978
   Detection Prevalence : 0.8524
      Balanced Accuracy : 0.8444

       'Positive' Class : 0
```

ROC Curve – AUC = 0.943



Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

DSM Evaluation

Discussion

Based on the DSM what would your decision/recommendation be? Why?

Both of these approaches yield good accuracy metrics. The accuracy metrics are ~97 % on the validation data sets, and the sensitivity and specificity for both are approximately 98% and 70%. The specificity for both models would predict more flights being cancelled than in reality. It is advised that this approach (over predicting rather than under) is a preferable conservative approach. The commendation is to use the logistic regression model due to its slightly higher accuracy metrics and AUC.

What are the limitations of the DSM you have used?

There are two primary weaknesses of our models. The first is that the training data is a historical account, and any output would be based on the presumption of stationarity. That is, the data must be "relevant" to current times. Right now, airlines are suffering and unprecedented impact to their bookings due to effects of COVID-19. Many airlines are running skeleton schedules at the moment, and the result is that DOT on-time performance is abysmal (due to many cancelled flights). But the operational flights have stellar OTP. That is, the flights that are still being flown are extremely likely to be on-time. Our models cannot be expected to be relevant in the current environment.

Another weakness is related to weather data. We were able to acquire data for wind speed, but not wind direction. It is easy for pilots to take off and land with relatively straight head or tailwinds. However, crosswinds can hinder whether or not a plane is able to safely take off or land. The presence of these cross winds can have a large impact on OTP. In order to gain insight to this feature, we would need the angle/direction of the runways being used for flights, and the direction of the winds.

What would you expect (most likely) to influence the decision-making process? How does the decision support mitigate some/all of these?

In pre-COVID19 times, I would expect the flight operations managers to weight the business KPI's and the revenue metrics for individual flights near-in to the departure dates (i.e. the week or day-of). The time of day will also have a big impact on flight operations. Late or on-time flights will have a propagation effect. Using historical data, were are able to successfully predict whether a flight will be on-time or late, and operations teams would be able to utilize this model output.

In reality, COVID-19 has disrupted almost all aspects of operation, and I believe that costs and revenues are taking priority. OTP, at the moment, is not a big concern because of reduced capacity and demand.

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

What enhancements would you aim for to enable better decision support for this task?

I would like to search for data regarding the wind direction (and speed) in relation to the orientation of the runways being used. At this time, it is unclear if this data exists. Logistic regression and decision trees are also still very valuable due their relatively short training times. Re-training on a frequent basis would be essential to prevent model drift. It would also be beneficial to explore other classification models such as support vector machines, neural networks, and naïve Bayes. Some of these would necessitate the use of better hardware, however.

Group 5 Final Project
Julian, Vegesna, Thomas, Hernandez

Appendix A: SQL Query for Raw Data

```
SELECT

EXTRACT(MONTH FROM flightdate) AS "month"

, td_day_of_week(flightdate) AS DoW

, marketingairline

, origin

, dest

, totalseatcount

, nautmiles

, EXTRACT(HOUR FROM out_) AS hour_out

, EXTRACT(HOUR FROM in_) AS hour_in

, airtime

, blockvariance

, depvariance

, divertedflag

, depdelayflag

, internationalflag

, origintemp

, originwindspeed

, CAST(CASE WHEN originwindgust = '-' THEN 0 ELSE originwindgust END AS INTEGER) AS
originwindgust

, originvisibility

, originskycondition1

, CAST(CASE WHEN origincloudlevel1 = '-' THEN '40000' ELSE origincloudlevel1 END AS INTEGER)
AS origincloudlevel1
```

```
, desttemp

, destwindspeed

 ,CAST(CASE WHEN destwindgust = '-' THEN 0 ELSE destwindgust END AS INTEGER) AS
destwindgust

, destvisibility

, destskycondition1

, CAST(CASE WHEN destcloudlevel1 = '-' THEN '40000' ELSE destcloudlevel1 END AS INTEGER)
AS destcloudlevel1

, acttimeongateorigin - schtimeongateorigin AS timeongatevariance

, arrvariance

, CASE WHEN arrvariance > 0 THEN 1 ELSE 0 END AS late_flag

, CASE WHEN arrvariance > 14 THEN 1 ELSE 0 END AS DOT_late_flag

--mas.*,

/*flightdate

,td_day_of_week(flightdate) AS DoW

,marketingairline

,operatingairline

,flightno

,origin

,dest

,CASE WHEN origin < dest THEN origin||'-'||dest ELSE dest ||'-'||origin END AS market

,origintimezoneoffset

,desttimezoneoffset

,CAST(CASE WHEN nextdayflag = '1' THEN 1 ELSE 0 END AS INTEGER) AS nextdayflag

,totalseatcount
```

,premiumseatcount

,firstseatcount

,busseatcount

,ecoseatcount

,generalacft

,routing

,statmiles

,nautmiles

,depgate

,arrgate

,actualtailnumber

,out_

,off_

,on_

,in_

,airtime

,actualblocktime

,scheduledblocktime

,blockvariance

,taxiout

,taxiin

,depvariance

,arrvariance

,divertedflag

```
,cancelledflag

,taxiout30flag

,taxiout60flag

,taxiout90flag

,depdelayflag

,arrdelayflag

,blockzeroflag

,domesticflag

,internationalflag

,origintemp

,origindewpoint

,originwinddirection

,originwindspeed

,CAST(CASE WHEN originwindgust = '-' THEN 0 ELSE originwindgust END AS INTEGER) AS
originwindgust

,originvisibility

,originwxstring

,originskycondition1

,origincloudlevel1

,originskycondition2

,origincloudlevel2

,originskycondition3

,origincloudlevel3

,desttemp
```

,destdewpoint

,destwinddirection

,destwindspeed

,destwindgust

,destvisibility

,destwxstring

,destskycondition1

,destcloudlevel1

,destskycondition2

,destcloudlevel2

,destskycondition3

,destcloudlevel3

,schbufforigin

,actbufforigin

,schbuffdest

,actbuffdest

,schtimeongateorigin

,acttimeongateorigin

,schtimeongatedest

,acttimeongatedest

,ronflag

,scheduledgatedepartureddatetime_zulu

,scheduledgatedepartureddatetime_zulu + CAST(origintimezoneoffset - 4 AS INTERVAL HOUR) AS scheduledDepartureLocal

```
,scheduledgatearrivaldatetime_zulu

,scheduledgatearrivaldatetime_zulu + CAST(desttimezoneoffset - 4 AS INTERVAL HOUR) AS
scheduledDepartureLocal

,date_rec_added

,specificacft

,brakes_set_ts*/

FROM LAB_NP_MASFLIGHT.blk_masflight mas

WHERE 1=1

AND flightdate BETWEEN '2018-01-01' AND '2019-12-31'

AND marketingairline IN('AA', 'AS', 'B6', 'DL',  'F9', 'G4', 'HA', 'NK', 'SY', 'UA', 'WN')

AND origintimezoneoffset IS NOT NULL

AND desttimezoneoffset IS NOT NULL

AND out_ IS NOT NULL

AND off_ IS NOT NULL

AND on_ IS NOT NULL

AND in_ IS NOT NULL

AND airtime IS NOT NULL

AND actualblocktime IS NOT NULL

AND scheduledblocktime IS NOT NULL

AND ronflag IS NOT NULL

AND depgate IS NOT NULL

AND arrgate IS NOT NULL

AND acttimeongatedest IS NOT NULL

SAMPLE 200000
```