Flight Delays
Full Data Mining Analysis
IBM – Modeling Recommendation

Technical Report

October 21, 2022

Prepared By:
Group 7

Mr. Paul Gulbrandsen U14430215

Mrs. Emma Horton U66789164

Mr. Raj Jeshwant Kumar Bandari U08911814

Mr. Albert Cela U55737379

Mr. Joseph Prete U09228992

Prepared For:
Dr. Kiran Garimella
Data Mining Course ISM 6136.020
Fall 2022
University of South Florida
Muma College of Business

1 Background of the Problem

This section was written by Paul Gulbrandsen

Our team consists of five members formed to analyze and present a specific subset of data on behalf of the International Business Machines (IBM) corporation. IBM is the leading American computer manufacturer with multiple locations throughout the United States and abroad. Although their clientele spans the globe, IBM has incorporated a client-centric culture “designed to accelerate the delivery of value to their clients and partners”. (2021 IBM Annual Report)

Based on the 2021 IBM Annual Report, 70% of IBM’s annual revenue was generated from their software and consulting divisions. IBM updated its client engagement model to improve how they deliver value to the customer. Their client-first, problem solving approach is the reason existing customers continue to trust IBM employees and new customers hire IBM.

Travel is a large part of the IBM business plan and as such IBM has chosen to hire our team of subject matter experts (SMEs) to perform data mining analysis on the efficiency of flights, which enable their employees to conduct regular client engagements. Each year, IBM pays millions of dollars in employee travel and transportation costs. Most of these trips are affected by the aviation industry and the efficiency of multiple airline companies. Flight delays or even worse, cancellations, negatively affect IBM’s client-centric approach.

IBM desires to know which flights employees should select to reduce the possibility of having their travel delayed or canceled because with every delay or cancellation, IBM loses money in lost time for employees, missed meetings with clients, and additional lodging expenses. The ability to forecast flights with the least possibility of disruption will enable employees to select flights to support IBM’s client-centric approach. The idea is to minimize the negative impact to client-employee meetings as well as minimize the negative affect to employee morale caused by flight delays and cancellations.

The goal of this project is to use data mining analysis to build machine learning models to identify flights that won’t be delayed or canceled. The data will enable our team to compare delays and cancellations by all the major airline companies using variables such as origin and destination airports, scheduled departure and arrival times, months, and days of the week. The goal is to ensure minimal disruption to the IBM traveling employee, minimize instances of meeting cancellations, and reduce the potential occurrence of additional lodging and transportation costs.

2 Motivation for Solving the Problem

This section was written by Joseph Prete

Every year flight delays and cancellations cause billions of dollars of wasted productivity not only on the airline industry’s bottom line, but also to passengers and society. It is estimated that domestic flight delays put a $32.9 billion dent in the U.S.

economy, and more than half of that cost is borne by airline passengers. This means that nearly $16.7 billion dollars of lost productivity are shouldered by passengers and is calculated based on lost passenger time due to delays, cancellations and missed connections, as well as expenses for food and accommodations because of being away from home. (Berkeley Engineering, 2021) (Ball, 2010) There is also a hidden component of waste that is caused because of inefficient flights where customers discourage flying because of previous negative experiences which then have downstream effects on associated businesses and ultimately hinders overall total gross domestic product. Flight efficiency is a multifactorial issue with factors that range from unavoidable ones such as mechanical and weather related to manageable ones such as capacity production and policy implementation.

As of 10/10/2022, IBM with headquarters in Armonk, NY has 297,800 employees and annual revenues of 57.3 billion dollars. (Staff, 2022) By these metrics, IBM is a massive company with proximity to Newark Liberty International Airport (EWR) which is the busiest International Airport in the US. This puts IBM in a great position to gain on cost savings by increased utilization of flight efficiency for its massive workforce. Another opportunity for savings includes flight booking flexibility. According to TravelPerk, when booking corporate travel it is important to request if possible three different arrivals and departures, and by accepting this flexibility and discovering a wiggle room in pricing could lead to savings of near 50% (Castillo, 2022) Thus IBM with its large workforce and accompanying demand should utilize carriers that are efficient in preventing delays as well as ones that fly frequently so the cheapest dates of travel can be booked and costs saved.

The reasons stated above were the motivating factors in seeking answers to our questions such as: Which carriers fly frequently efficient flights? Which days are the best to book? Which are carriers to avoid based on cancellations? By answering these questions IBM will make better business decisions when booking corporate travel and ultimately lead to cost savings and corporate growth.

3 Solution Methodology and Evaluation Metrics

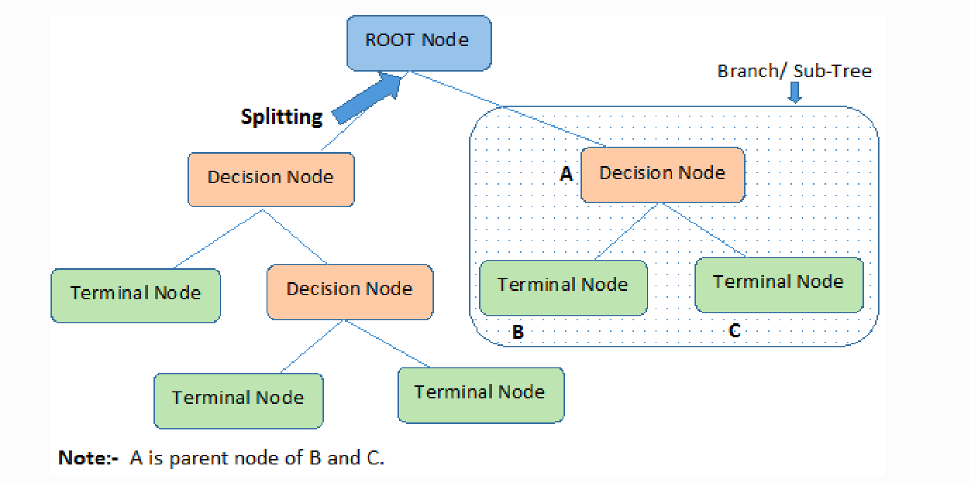
This section was written by Raj Jeshwant Kumar Bandari

Flight Delays are a significant issue in the aviation sector. Over the past two decades, major expansion in the airline industry has resulted in air traffic congestion, resulting in flight delays. In general, a flight is considered delayed when takeoff deviates from the scheduled time resulting in a delayed arrival. The airline data collected maintains a variable that indicates when the arrival time of the flight and the scheduled arrival time differ by 15 minutes or more.

There are several methodologies that can be used to perform the data mining techniques. In this project, we are using a Two-Class Boosted Decision Tree and a Two-Class Logistic Regression model. Logistic regression is a popular machine learning technique. Its primary advantages are the clarity of the results and ability to illustrate the relationship between dependent and independent features in a clear manner. It uses less processing power and is generally faster than Random Forest or Gradient Boosting.

Logistic regression was chosen to model flight delays and cancellations for a variety of reasons. First, the weights of each feature trained by logistic regression are easily interpretable, as the sign of the weight indicates if a flight is more or less likely to be delayed if it has a high value for that feature. Second, logistic regression outputs provide a measure of confidence through the probability of belonging to each class.

Decision Trees are used as a decision-taking tool which uses a flowsheet like tree formation. Decision Tree can also be defined as representation of decisions and all the possible outcomes. Decision Trees can be utilized for both continuous and categorical outputs and the conditions are illustrated by the Decision nodes and results by the end nodes.



Note:- A is parent node of B and C.

4 Description of Dataset

This section was written by Raj Jeshwant Kumar Bandari

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2013 flight delays and cancellations.

Each entry corresponds to a flight, and we can see that over 135,000 flights were recorded in 2013. These flights are described by 17 parameters (such as Date, Flight number, Departure, Arrival, etc.). A description of these variables can be found below:

1. **Year**: The year of the flight (all records are from 2013)
2. **Month**: The month of the flight
3. **DayofMonth**: The day of the month on which the flight departed
4. **DayOfWeek**: The day of the week on which the flight departed - from 1 (Monday) to 7 (Sunday)
5. **Carrier**: The two-letter abbreviation for the airline.

6. **OriginAirportID**: A unique numeric identifier for the departure airport
7. **DestAirportID**: A unique numeric identifier for the destination airport
8. **CRSDepTime**: The scheduled departure time
9. **CRSArrTime**: The scheduled arrival time
10. **ArrDelay15**: A binary indicator that arrival was delayed by more than 15 minutes (and therefore considered "late")
11. **Canceled**: A binary indicator that the flight was canceled
12. **OriginCity**: The departure airport city
13. **OriginState**: The departure airport state
14. **OriginName:** The full name of the departure airport
15. **DestCity**: The destination airport city
16. **DestState**: The destination airport state
17. **DestName:** The full name of the destination airport

When utilizing the data in our models, we choose to focus on key variables. The variables selected for the models were: Arrival Delays greater than 15 minutes, Cancellations, Month, Day of the Week, Carrier, Departure Time, Origin State, and Destination State. For the purposes of the client requirements and to tune for the best performance, multiple models were trained using Cancellations and Delays greater than 15 minutes. This gave us the ability to analyze the data to help select flights that minimize the negative effects of both possibilities.

5 Comparison of Algorithms

This section was written by Emma Horton

The two algorithms selected for the analysis were Two-Class Boosted Decision Trees and Two-Class Logistic Regression models due to their ease of use and robust nature. Both algorithms provide the ability to perform effectively on a variety of machine learning tasks when properly configured, predict an outcome variable in a supervised format, and are optimized for binary variables such as cancelled flight or not cancelled and delayed flight or not delayed.

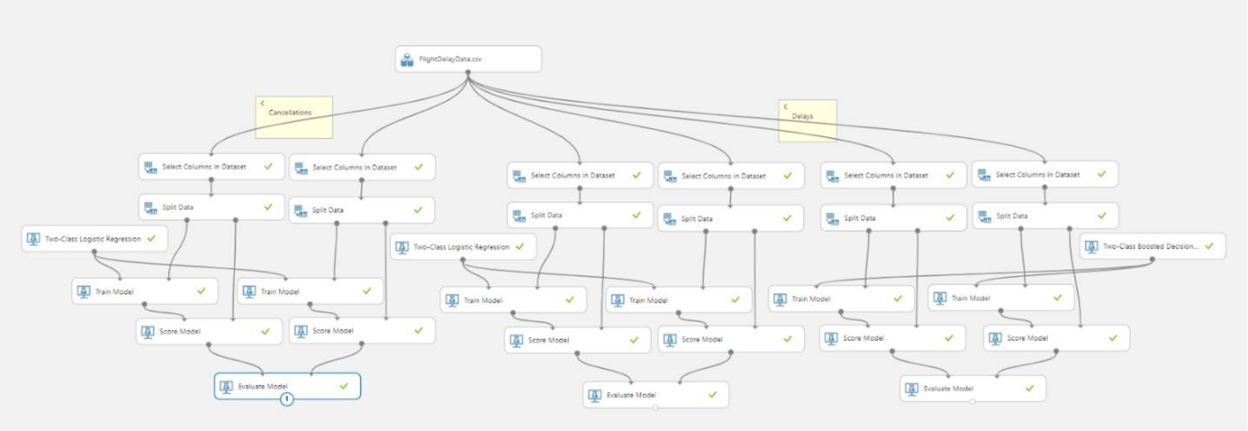
A Boosted Decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the second and so forth. A drawback of decision trees would be how memory-intensive it is, so it is better for smaller datasets. The settings for a decision tree include single parameter or range, maximum number of leaves per tree (the value can improve precision but risk overfitting when increased and initiates a longer train time), minimum number of samples per leaf node indicates the number of cases required to create any terminal node (with a default value of 1, a single case can create a rule), learning rate defines the step size while learning and it determines how fast or slow the learner converges on the optimal solution (step size too large learner will overshoot solution, step size too small training takes longer to converge on the best solution), number of trees constructed is the total number of trees to create in the ensemble (more trees can mean better coverage but increased training time) and finally random number seed ensures reproducibility cross runs with same data and parameters.

A Two-Class Logistic Regression is a method to predict the outcome and is popular in classification use-cases. The algorithm predicts the probability of occurrence of an event by fitting data to a logistic function. This is a supervised learning method where an outcome variable is used to train the model, in this case Cancellations or Arrival Time Delay. The configuration for this algorithm includes single parameter or parameter range, optimization tolerance (which specifies a threshold value to use when optimizing), if the improvement between iterations falls below threshold the solution is reached and training stops, L1 and L2 regularization weight values are typically non-zero and is a method to prevent overfitting by penalizing models with extreme coefficient values by adding the penalty to the error of the hypothesis (meaning an accurate model with extreme coefficient values would be penalized more), and random number seed is an integer value and makes results reproducible over multiple runs.

6 Summary of Experiments

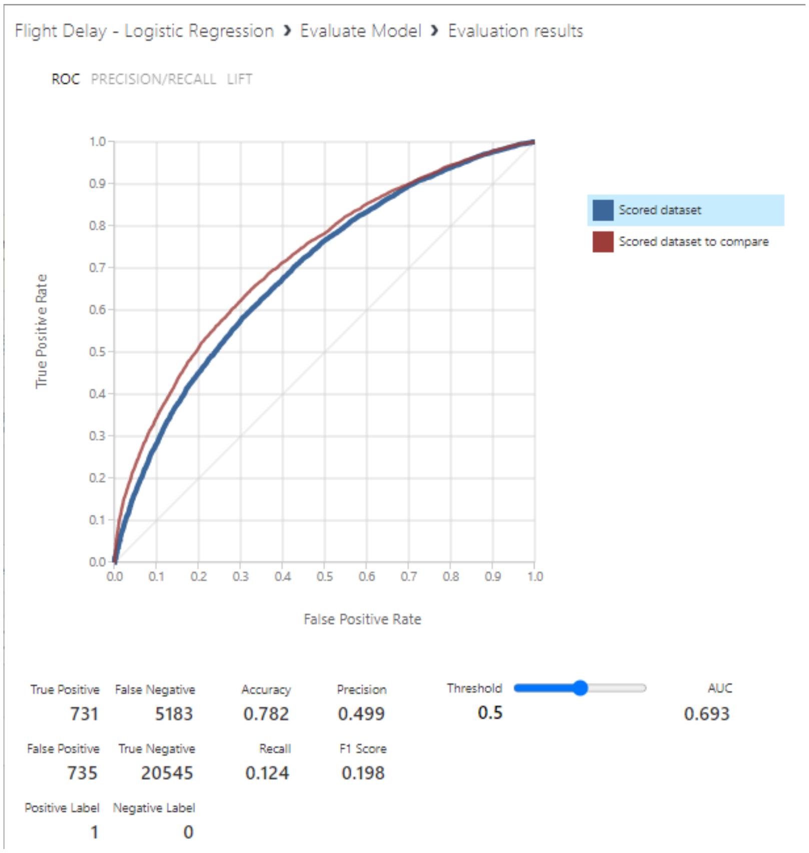
This section was written by Emma Horton and Paul Gulbrandsen

The primary focus of the experiment was on training the data using delays greater than or less than 15 minutes. The data science team proposed running two algorithms to test for the best performance. Per the parameters of the business need, a decision tree and a logistic regression were selected for use. Two tests were performed for each algorithm to identify the best criteria for the model. In addition, the team also conducted two Logistic Regression models using cancellations to train the data. Below is a visual of the model produced. This holistic model was produced by the team lead, **Paul Gulbrandsen**, but other members of the team ran one individual experiment and sent the results to the team lead for consolidation in this paper. Team member names are included with the description of the model the member created.



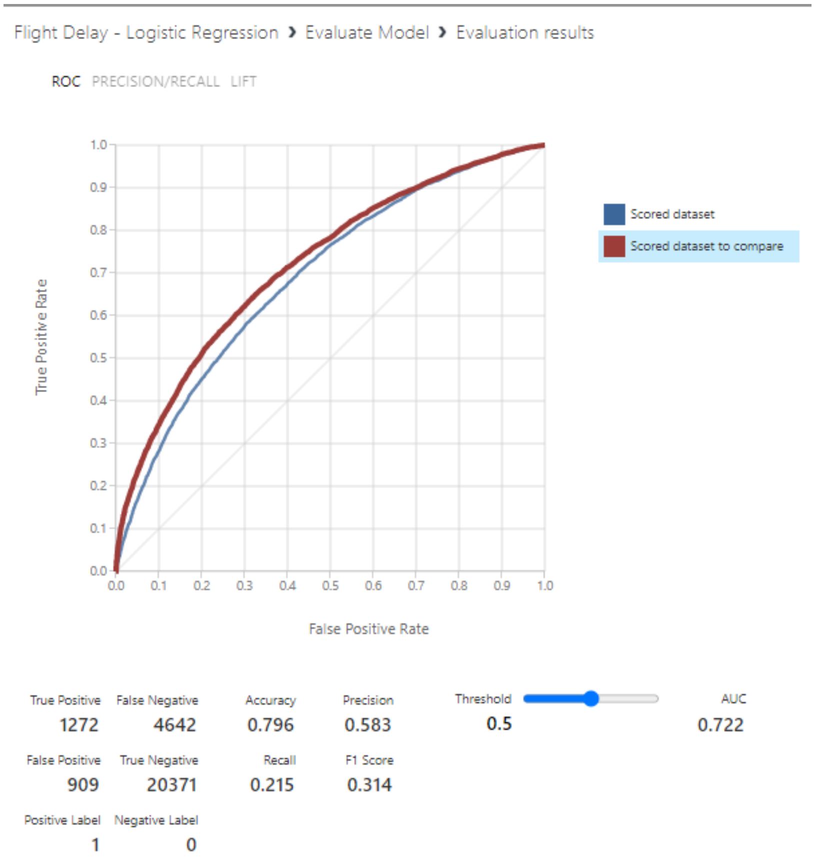
Decision Tree Experiments:

The first test was performed by **Emma Horton** and used a decision tree algorithm in the Microsoft Azure cloud resource. The data set was limited to Month, Day of the Week, Carrier, Departure Time, Origin State, Destination State, and Arrival Delay greater than 15 minutes. The Decision tree was run with a .80 fraction of rows in the first output dataset, with a Maximum number of leaves per tree of 20 and Minimum number of samples per leaf node of 10, a learning rate of .2, and number of trees constructed was set to 100. The model was trained on the variable ArrDel15 which is a binary variable indicating whether the flight was delayed more than 15 minutes. The model performed as follows with an Accuracy of 78.2%, a Precision of 49.9%, a Recall of 12.4%, an F1 Score of 19.8%, and a **Specificity of 96.5%**.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
731	5183	0.782	0.499	0.5	0.693
False Positive	True Negative	Recall	F1 Score		
735	20545	0.124	0.198		
Positive Label	Negative Label				
1	0				

The second Decision tree was performed by **Paul Gulbrandsen**. This experiment added Origin Airport ID, Destination Airport ID, Day of the Month, Origin City and Destination City. The algorithm was ran using .80 fraction of rows with the same criteria in the first; Maximum number of leaves per tree of 20 and Minimum number of samples per leaf node of 10, a learning rate of .2, and number of trees constructed to 100. The model was also trained on the variable ArrDel15. The model performed with very similar results with an Accuracy of 79.6%, a Precision of 58.3%, a Recall of 21.5%, an F1 Score of 31.4%, and **Specificity of 95.7%**.



Flight Delay - Logistic Regression › Evaluate Model › Evaluation results

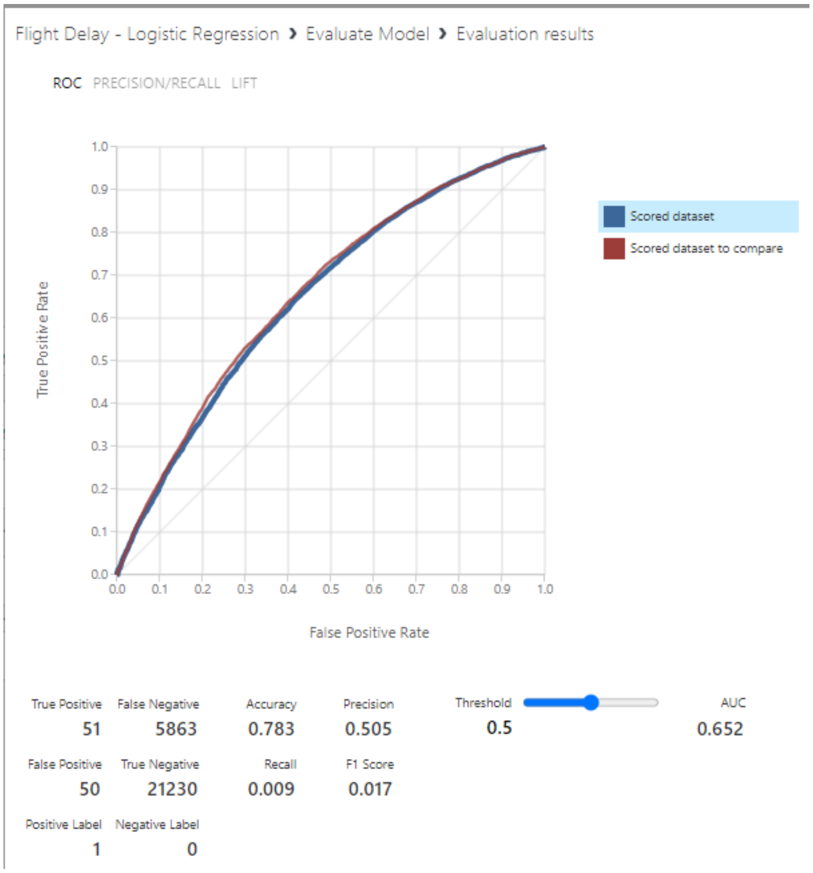
True Positive 1272	False Negative 4642	Accuracy 0.796	Precision 0.583	Threshold 0.5	AUC 0.722
False Positive 909	True Negative 20371	Recall 0.215	F1 Score 0.314		
Positive Label 1	Negative Label 0				

The curve illustrates the reliability of the models depending on the prediction threshold. The larger the area under the curve, the better the prediction model. Based on our results and the slight changes seen, we conducted an analysis of the data set in greater detail and noticed the dataset is unbalanced. Meaning that 78% of the data reflects flights that are not delayed greater than 15 minutes and only 22% of the flights were delayed greater than 15 minutes. Based on this analysis the importance of the Accuracy and Precision results were not given as much consideration in our recommendation as the Specificity. When a data set is unbalanced, specificity is the best metric to use. In both decision tree tests; the specificity was greater than 95%. This is a good indicator of the models' ability to identify flights that won't be delayed.

Logistic Regression Experiments:

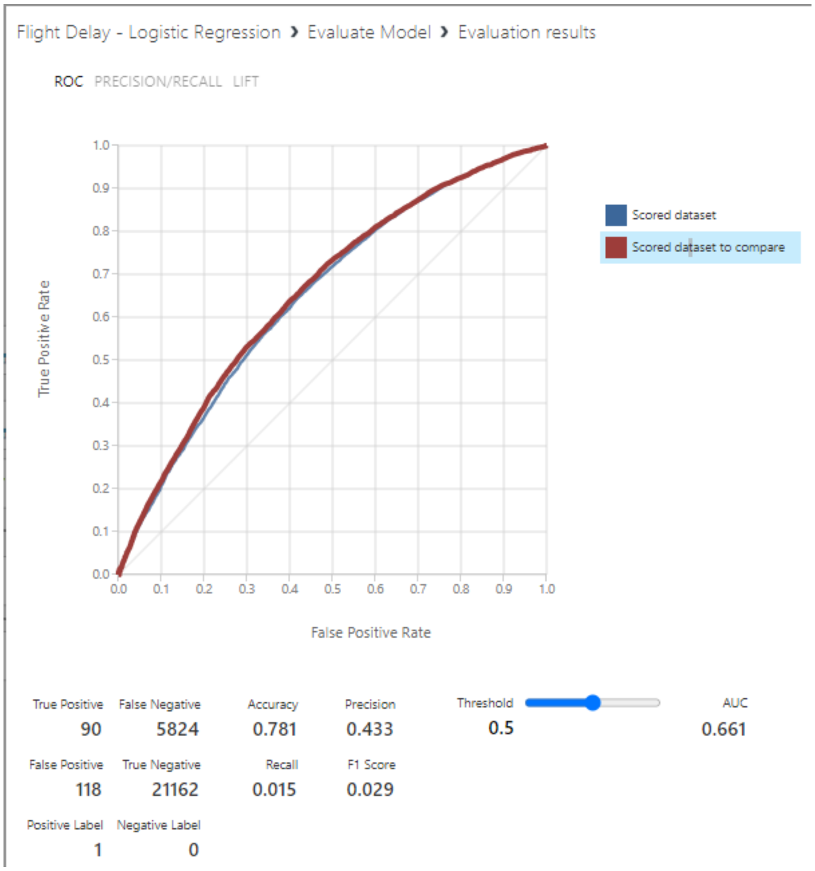
The first logistic regression was run by Emma Horton using an optimization tolerance of 1E-07, an L1 regularization weight of 1, an L2 regularization weight of 1, and a memory size for L-BFGS of 20. Like the first Decision Tree, the data set was limited to Month, Day of the Week, Carrier, Departure Time, Origin State, Destination State, and Arrival Delay greater than 15 minutes. The fraction of rows utilized mirrored that of the Decision tree with a .80 fraction of rows in the first output dataset. The model was also trained on the variable ArrDel15 which is a binary variable indicating whether the flight was delayed more than 15 minutes. The model performed as follows with an Accuracy of 78.3%, a

Precision of 50.5%, a Recall of 0.9%, an F1 Score of 1.7%, and a **Specificity of 99.8%**. The ROC Curve is sloped upwards and is above the gray line, which means that the model works better than random assumptions. The gray line corresponds with a 50% chance to accurately predict results.


Flight Delay - Logistic Regression › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
51	5863	0.783	0.505	0.5	0.652
False Positive	**True Negative**	**Recall**	**F1 Score**		
50	21230	0.009	0.017		
Positive Label	**Negative Label**				
1	0				

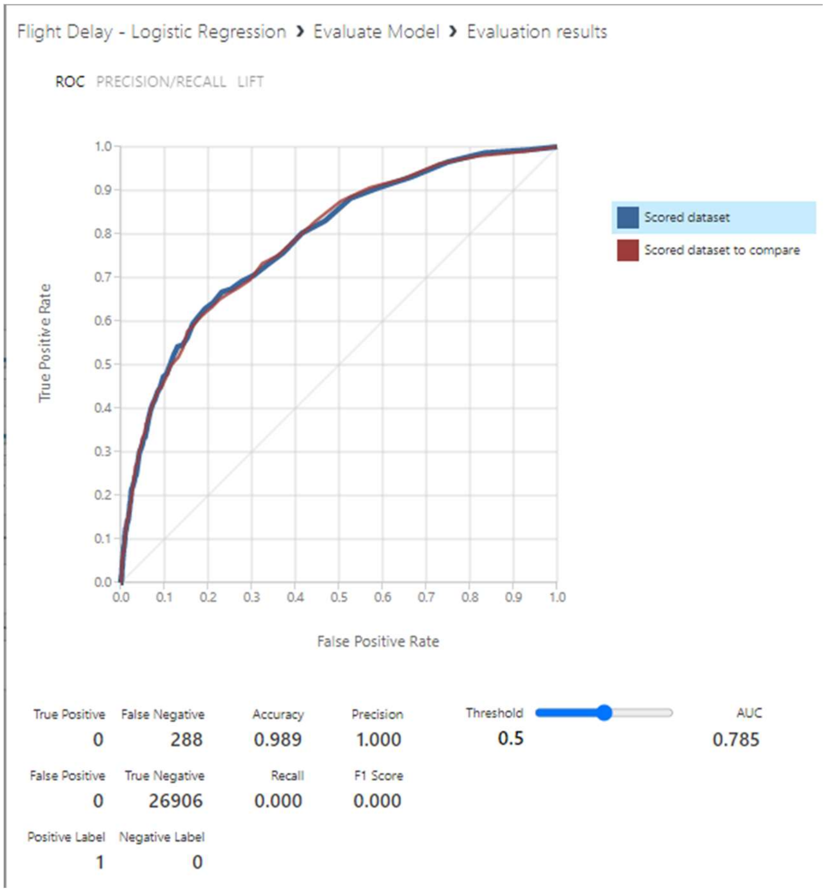
The second logistic regression model was performed by **Paul Gulbrandsen**. This model mirrored the selected columns used in the second Decision Tree. Therefore, it contained: Month, Day of the Week, Carrier, Departure Time, Origin State, Destination State, Origin Airport ID, Destination Airport ID, Day of the Month, Origin City, Destination City, and Arrival Delay greater than 15 minutes. The model was run using the same criteria as the first logistic regression model with an optimization tolerance of 1E-07, an L1 regularization weight of 1, an L2 regularization weight of 1, and a memory size for L-BFGS of 20. The model was trained on the variable ArrDel15 which is a binary variable indicating whether the flight was delayed more than 15 minutes. The model performed as follows with an Accuracy of 78.1%, a Precision of 43.3%, a Recall of 1.5%, an F1 Score of 2.9%, and a **Specificity of 99.4%**. The ROC Curve is sloped upwards and is above the gray line, which means that the model works better than random assumptions.



Flight Delay - Logistic Regression › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
90	5824	0.781	0.433	0.5	0.661
False Positive	True Negative	Recall	F1 Score		
118	21162	0.015	0.029		
Positive Label	Negative Label				
1	0				

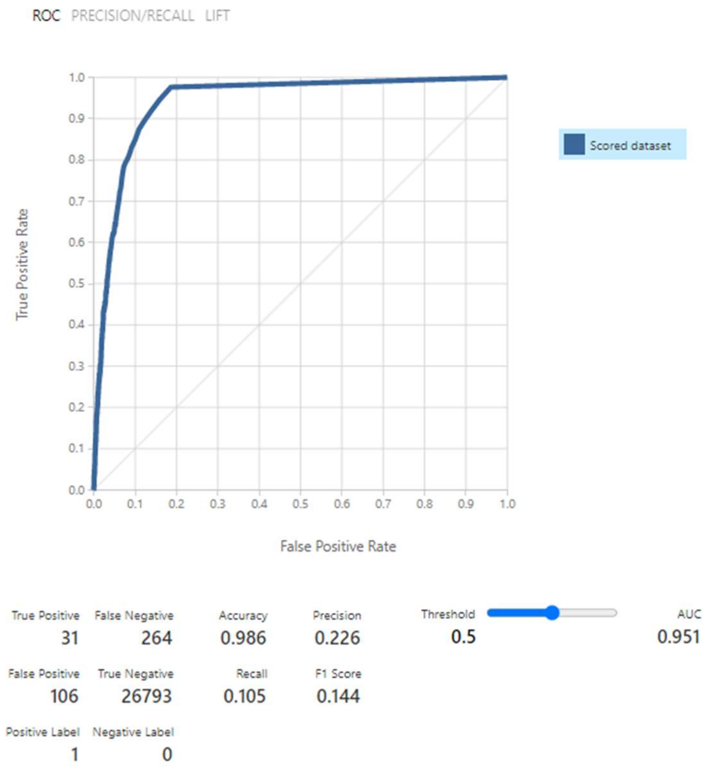
In addition to running models on flight delays, we also wanted to look at cancellations. **Joseph Prete** took the same criteria for the two Logistic Regression models and replaced Flight Delays greater than 15 minutes with Cancellations. Additionally, since flight delays were removed, these two models were trained using cancellations, which is also a binary variable. The results for the two models were identical: Accuracy of 98.9%, Precision of 100%, Recall of 0%, F1 Score of 0%, and Specificity of 100%. Once again, since these results appeared skewed an analysis of the data set was conducted and we discovered that the data was even more unbalanced with 99% of the data reflecting flights that were not canceled and only 1% where flights were canceled. Only one graphic is shown since the results are identical. However, you can see that both the red and blue dataset lines are present on the graph.



Flight Delay - Logistic Regression › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
0	288	0.989	1.000	0.5	0.785
False Positive	**True Negative**	**Recall**	**F1 Score**		
0	26906	0.000	0.000		
Positive Label	**Negative Label**				
1	0				

Albert Cela looked at cancellations from a Decision tree model standpoint. The same conditions as in the first run of the algorithm were applied: a learning rate of.2, a maximum number of leaves per tree of 20, a minimum number of samples per leaf node of 10, and a number of trees generated to 100. The variable Cancelled was used to train the model. The model accurately predicted that 31 flights would have been canceled (true positives). In 106 situations, the model was inaccurate (false positives). The model predicted that 264 were canceled, and that came to fruition (false negatives). In 26,793 instances, the model's prediction that these flights would not have been canceled was accurate. In general, 98.9% of the time the model is right (Accuracy = 0.801). We then look at the ROC curve. Depending on the prediction threshold, the curve shows how reliable the model is. The accuracy of the prediction model increases with the area under the curve. The upward-sloping curve and the fact that it is above the gray line indicate that the model performs better than arbitrary assumptions. The gray diagonal line represents a 50% possibility of lying truthfully, making it simple to estimate. The area would be 1.0 with a flawless exact model for each flight. This Azure Machine Learning Studio-created prediction model can predict with 98.9% accuracy whether flights on routes will be canceled or not.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
31	264	0.986	0.226	0.5	0.951
False Positive	**True Negative**	**Recall**	**F1 Score**		
106	26793	0.105	0.144		
Positive Label	**Negative Label**				
1	0				

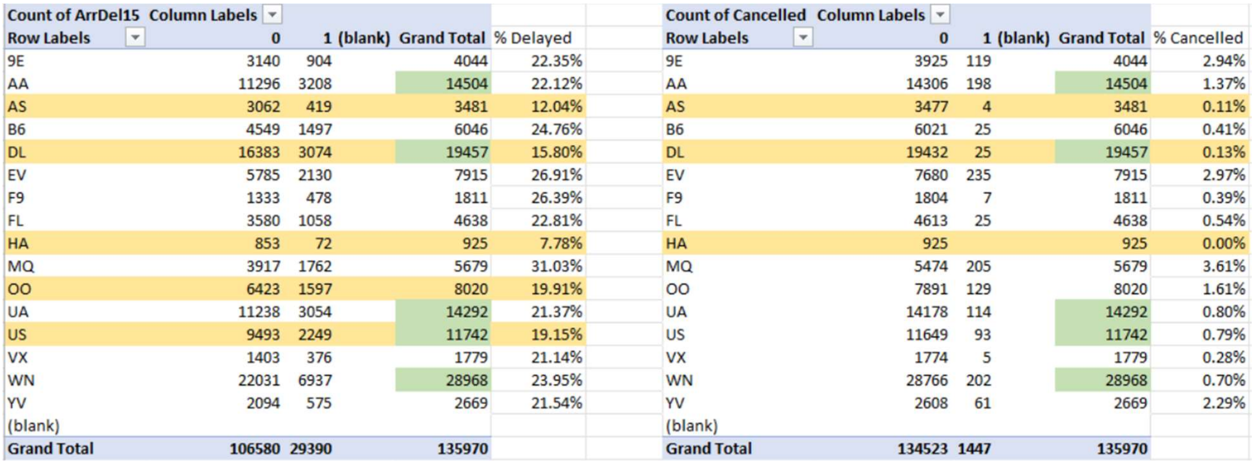
7 Conclusions and Recommendations

This section was written by Albert Cela and Paul Gulbrandsen

In Azure Machine Learning Studio, we produced multiple models in order to predict flight delays and canceled flights. From our efforts we were able to predict with certain confidence that our model will be able to identify flights that won't be delayed. Since the data set was unbalanced, we used Specificity to justify our recommendation. In the case of the delayed flights both Decision Trees provided a specificity of about 96%. Meaning that an IBM employee would be able to select a flight with almost 96% certainty that the flight would not be delayed. In the Logistic Regression models the Specificity for delayed flights was even greater with both producing results of 99%. Therefore, using the combination of both the Decision Tree and the Logistic Regression models we are very confident in our model’s ability to allow IBM employees to select a flight that will not be delayed and cause additional lost time for employees, missed meeting with clients, or additional travel and lodging expenses for IBM.

When we take into consideration the models showing cancellation data, our models reflect that an individual would be able to identify with 100% certainty, those flights which will be canceled using Month, Day of the Week, Carrier, Departure Time, Origin State, Destination State, Origin Airport ID, Destination Airport ID, Day of the Month, Origin City, and Destination City as the criteria. However, since the data set only contained data on about 1200 canceled flights out of a total of about 108,700 flights, it would be prudent to caveat our recommendation to reflect that the model may be skewed.

In addition to using the models we analyzed the data set using pivot tables and were able to make recommendations to reflect which carrier the IBM employees should look to first when seeking out flights. The pivot table on the left reflects delayed flights by carrier and the pivot table on the right reflects canceled flights by carrier. On the left the yellow highlight shows carriers with less than 20% flight delays. The green highlight shows carriers with more than 10,000 flights. On the right the yellow highlights show carriers with less than 0.2% cancellations. The green highlights show the carriers with greater than 10,000 flights. Based on the data it would be best for employees to utilize our model by first looking at flights run by DL since they are one of the largest airlines with the least amount of cancellations and delays.



Count of ArrDel15	Column Labels				
Row Labels	0	1	(blank)	Grand Total	% Delayed
9E	3140	904		4044	22.35%
AA	11296	3208		14504	22.12%
AS	3062	419		3481	12.04%
B6	4549	1497		6046	24.76%
DL	16383	3074		19457	15.80%
EV	5785	2130		7915	26.91%
F9	1333	478		1811	26.39%
FL	3580	1058		4638	22.81%
HA	853	72		925	7.78%
MQ	3917	1762		5679	31.03%
OO	6423	1597		8020	19.91%
UA	11238	3054		14292	21.37%
US	9493	2249		11742	19.15%
VX	1403	376		1779	21.14%
WN	22031	6937		28968	23.95%
YV	2094	575		2669	21.54%
(blank)					
Grand Total	106580	29390		135970	

Count of Cancelled	Column Labels				
Row Labels	0	1	(blank)	Grand Total	% Cancelled
9E	3925	119		4044	2.94%
AA	14306	198		14504	1.37%
AS	3477	4		3481	0.11%
B6	6021	25		6046	0.41%
DL	19432	25		19457	0.13%
EV	7680	235		7915	2.97%
F9	1804	7		1811	0.39%
FL	4613	25		4638	0.54%
HA	925			925	0.00%
MQ	5474	205		5679	3.61%
OO	7891	129		8020	1.61%
UA	14178	114		14292	0.80%
US	11649	93		11742	0.79%
VX	1774	5		1779	0.28%
WN	28766	202		28968	0.70%
YV	2608	61		2669	2.29%
(blank)					
Grand Total	134523	1447		135970	

8 References

1) Berkeley Engineering. (2021, July 27). Flight delays cost more than just time. Retrieved October 20, 2022, from https://engineering.berkeley.edu/news/2010/11/flight-delays-cost-more-than-just-time/

2) Ball, M., Barnhart, C., Dresner, M., & Hansen, M. (2010). *Total Delay Impact Study*. Nextor. Retrieved October 22, 2022, from https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf

3) Staff, F. (2022, October 10). *IBM*. Fortune. Retrieved October 20, 2022, from https://fortune.com/company/ibm/

4) Castillo, À. (2022, October 20). *9 ways your company can save money on business travel*. TravelPerk. Retrieved October 20, 2022, from https://www.travelperk.com/blog/ways-your-business-can-save-money-on-business-travel/