An Analysis of Houston Flight Delays

Nigel Cox*, Meghan McGee*, Lucas Stefan*, Emily Veenstra*
*Department of Electrical and Computer Engineering, University of Calgary, Canada
{ntcox, meghan.mcgee, lstefan, ecveenst} @ucalgary.ca

***Abstract*—Flight delays and cancellations can result in air- and ground- traffic congestion, cost implications to travelers and airlines, as well as propagate further delays in flight patterns. Creating a model to predict when delays and cancellations happen, as well as to classify why they are happening, would allow travelers, airlines, and airports to plan for changes and minimize repercussions. The analysis in this paper is performed on flights leaving from George Bush Intercontinental Airport in Houston, Texas. Classification and multiple linear regression are used to predict when a delay or cancellation may happen, and how long the delay will be. Additionally, a model will be created to determine how different variables, or combinations of variables, affect the type of delay or cancellation. A K-nearest neighbor classifier will be built to predict whether a flight will be cancelled, delayed, or diverted; logistical regression and random forest classification will be used to predict the reason for the delay; and linear regression will be used to predict how long a delay may be. The aim of this study is to build models that will allow travelers to more confidently select reliable flights, and allow airlines and airports to gain insight on changes to flight patterns that need to be adjusted for, as well as operational inefficiencies.**

***Index Terms*—machine learning, flight prediction, classification, regression**

I. Introduction

Last year in the United States, over one million flights (approximately 20% of total departures) were delayed with another one hundred thousand cancelled [1]. The causes for these delays and cancellations vary and have been classified by the Department of Transportation as either Security, Weather, Carrier, National Air Service (NAS), and Late Arrival Delays [2]. A study performed by The Federal Aviation Administration (FAA), alongside research group Nextor, estimated that the annual cost of delays was approximately $26.6 Billion in the United States alone [3]. Another study by the FAA back in 2010 estimated that over half of the total cost of delays is paid by the passengers themselves [4]. Travelling can be a stressful and expensive experience for passengers, even without taking the possibility of delays or cancellations into account.

The Bureau of Transportation Statistics has made American domestic flight information from the year 1987 accessible to the public. With this data-set the main purpose of this project is to determine delay or cancellation trends based on a variety of different factors. Using the results of this analysis, a predictor will be created to allow passengers to more confidently select flights with less fear of delays or cancellation. Further analysis will also be performed to predict the length of delays with linear regression. As previously mentioned, there are different types of delays/cancellations; two models using logistic regression and random forest classification will be designed and tested to see how accurately the type of delay/cancellation can be predicted based on numerous variables.

The remainder of this paper is structured in the following format: Section II discusses related work that has previously been done and their findings. Section III describes the Research Questions this project is aiming to answer as well as the process of data collection, preparation, and analysis. Section IV discusses the results from the data-set while Section V goes over the key findings. Lastly, Section VI covers conclusions.

II. Related Work

The paper "A Review on Flight Delay Prediction" looks at the different applications and methodologies that are currently being used to predict delay and cancellation patterns in flights. The authors noticed that in recent years, there has been an increase in the number of papers written on flight delay analysis, specifically those using machine learning. The paper focuses on predicting propagation delays, and how flights, airlines, and airports may be affected by one major delay. Some of the techniques that are currently used to predict delay propagation patterns include statistical analysis and regression modes, as well as K-nearest neighbour, neural networks, fuzzy logic, and random forests. These techniques predict root delays which can propagate further delays, as well as predict delays relating to specific airport or airline interference such as taxiing in-and-out air traffic [5]. This paper reveals the importance of machine learning in efficient air transportation traffic, and provides insight to how travellers, airlines, and airports can save money and further propagated delays when using data to streamline air- and ground traffic.

Research has also been completed to evaluate the performance of an aircraft based on their flight metrics including: time to destination, distance traveled, delays encountered, and external conditions regarding the flight [6]. Evaluating the performance of an aircraft can help reduce air and ground delays, resulting in a safer airspace traffic control. Modeling and analyzing the performance of air carriers to more efficiently plan maintenance and flight patterns can result in huge savings in terms of environmental impact, fuel consumption, airline business models, and airspace optimization [1]. The paper Machine Learning Model for Aircraft Performances used attributes such as company, arrival, departure, day/hour, etc to predict the performance of an aircraft and predict possible delays. These attributes are used in a multidimensional classification model to predict the flight performance. To further refine the performance predictor implemented in the paper above, unsupervised methods for association rules

are implemented and flights that are most similar to the flight being analyzed are found. This final step is used to obtain the best prediction of aircraft performance based on historical data.

Zonglei et al developed a new model to alarm the approach of large-scale flight delays [7]. They initially used unsupervised learning methods, namely clustering, to identify different classes of delay. This method allowed them to label a data set into different classes of delay. Then, supervised learning methods were used to create an alarm model which can send an alert indicating the expected class of upcoming delays. Four separate models were created using Naïve Bayes, decision tree, neural network, and Ridor rule model methods. The decision tree method performed best; it could predict delays with an 80% confidence level.

III. Methodology

A. Research Questions

1) *How accurately can a flight delay, diversion, or cancellation be predicted?*

   This research question will involve attempting to predict flight delays, diversions, and cancellations for flights leaving from the George Bush Intercontinental Airport in Houston. Being able to predict future delays or cancellations would be helpful for travellers. They could choose flights which are more likely to be on time or could prepare a back up plan if the flight is likely to be affected. Additionally, this analysis will allow airports to see when they need to better prepare for commonly delayed or cancelled flights, while also gaining insight to inefficiencies in their operations.

2) *How do the different types of delays and cancellations depend on the variables surrounding an arrival or departure?*

   This question will discover what variables and combinations of variables affect the type of flight delay. The different types of delays recorded in the data set are carrier delays, weather delays, National Air System delays, security delays, and late aircraft delays. Any trends found in this investigation can help airlines or airports discover what sets of circumstances lead to different types of delays. These companies could then improve the reliability of their service. This would increase customer confidence as well as allow airline companies and airports to improve their business model. Also, less delays makes flying a better experience for everyone.

3) *Can the length of the delay be calculated before a delay happens?*

   The goal of this research question is to determine how accurate a prediction can be generated for the length of the delay. The analysis will be done on two main cases. The first is without any information about the type of delay. These results can be used at the outset to better understand the time consumption of processes that an airport uses to handle the different types of delays. The second is with information about the type of delay, representing the scenario when an airport is alerted to a delay and wants to know how long it will take to resolve. This can be useful in planning for a certain delay in order to reduce delay times at the airport, which can lead to a smoother operation.

B. Data Collection and Preparation

Approximately 32 GB of flight data was collected from the Bureau of Transportation services [1]. The data is stored in multiple csv files. These files contain 44 columns of data including information about airlines, arrival and departure airports, delay types and times, cancellation types and times, flight time, etc.

To prepare the data for analysis in the research questions, the data was manipulated using Spark data frames. The following are the pre-processing steps that were performed:

- **Collect data:** The data was downloaded from online to the ARC cluster at the University of Calgary where it could be processed in parallel using Spark.
- **Explore data:** The data was strongly structured so some time was spent becoming familiar with this structure. For example, the legend had to be studied to understand that CRS_ARR_TIME is the scheduled arrival time and ARR_TIME is the actual arrival time.
- **Visualize Data:** Visual distributions of the data were created to provide insight into what could be interesting to investigate and what could cause problems in our analysis.
- **Filter Data:** The filter() function of Spark data frames was used to remove data irrelevant to the research question. Irrelevant features were removed such as the taxi out time. Any flights that did not leave from the George Bush Intercontinental Airport (IAH) were removed. Some research questions required all flights and others required just delayed flights, so those were filtered accordingly.

  Additionally, the data was filtered to only the latest years. This was done to reduce the size of the data set. The machine learning was done on a single computer so the data had to be reduced to a size that could fit on a single machine's RAM.
- **Label Data:** Using Spark's withColumn() function, a new column was added to the data, which contained the label to be used in the supervised machine learning analysis. The labelling was done automatically by converting existing columns into an acceptable format. For example: on time = 0, delayed = 1, diverted = 2, and cancelled = 3.

- **Missing data:** The corpus does contain missing data. Since the data set is strongly structured, when large amounts of data are missing there is generally a logical reason. For example, if the arrival data is missing, this is because the flight was diverted or cancelled. This was managed by adding the diverted label. There were a small number of data points with null values for no obvious reason. These data points were removed. Delay type labels were missing from the early data (prior to 2003). For this reason, the analysis was limited to later years.

Spark is used to parallelize the processing of large amounts of data on multiple machines. Tests were run to see the effect that using different numbers of nodes has on the speed of pre-processing as shown in figure 1.

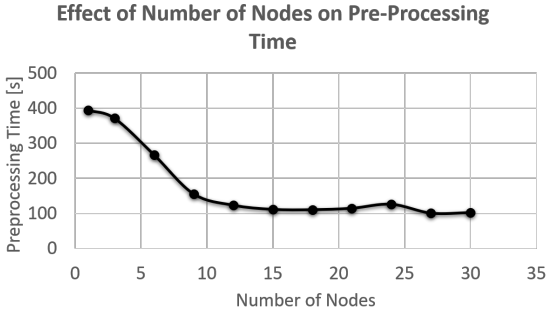

Fig. 1. Effect of the number of nodes on speed of the pre-processing for research question 1.

Adding more nodes increases processing speed, as expected. However, at about 13 nodes, the speed does not improve with more nodes. This is due to bottlenecks in the process that cannot be parallelized, such as loading the csv file into the Spark framework.

C. Data Analysis Plan

1) KNN Classification will be employed to determine how accurately flight delays, diversions, and cancellations can be predicted. After data collection and cleaning, a subset of three years worth of flight data (2010-2012) from the George Bush Intercontinental Airport (IAH) will be used for analysis. The first step will be to randomly split the data into a 80% - 20% training and test sets based on the X-variables and y-output (on-time, delayed, diverted, cancelled). As many of the X-variables are categorical, such as airline or destination airport, these have to be converted to numerical data to be analyzed with KNN. This will be done with the use of an encoder to create dummy variables for these categorical columns.

   Once the data has been properly converted from its categorical form, the first KNN test will be performed by selecting a k value equal to the square root of the number of training data points. From there the results will be analyzed by collecting the test set score and comparing the predicted values to the actual with the use of a confusion matrix and classification report. This classification report displays important information such as the precision, recall, f1, and support scores. Based on the results, different k values will be selected to see if the model can be improved. As the vast majority of flights are on-time it is likely that the k value will have to be decreased to ensure that the model isn't underfit and predicting that all flights are on-time.

2) Logistic regression and random forest classification will be used to predict the type of a delay based on features such as airport, airline, time of year, etc. Through data investigation, it was determined that approximately half of the delayed flights have more than one delay-classification; however, the overall delay time corresponds to the delay type with the largest time impact. Because of this, one-vs-rest multi-label logistic regression was performed and compared to single-label classification that categorized the delay type based on the highest time impact per delay. For example, if a flight was delayed by 5 minutes due to weather, and delayed by 25 minutes due to a late aircraft arrival, it would be classified as "late aircraft arrival". Single-label classification was also completed using random forest classification to compare the two methodologies.

   After cleaning data for each analysis model, discrete features in the data-sets were encoded using one-hot-encoding. The data-sets were split randomly into 70% - 30% training and test sets. First, the multi-label one-vs-rest logistic regression was performed. This analysis evaluates each classification type independently, and allows flights to be classified into more than one category. The accuracy, precision, recall, and F1 score were evaluated for each classification type. Next, logistic regression was used to perform single-label classification based on the most prominent delay type in the training set. Random forest classification was used to determine if a more accurate prediction could be obtained for classifying the type of flight delay. Feature extraction based on the random forest classifier was performed to determine which categories had the most influence on flight delays. The analyses were re-run based on the most important features to minimize run-time and increase model performance.

3) In order to predict the length of a delay, multiple linear regression must be used since the output will need to be continuous. The first part of the analysis will include encoding the discrete columns into dummy variables and splitting the entire data set into testing and training

sets. The important thing to note is that there will be three data sets to start with. One being a data set with all the information about the type of delay, and the other two having no information about the type of delay; one with all flights, and the second only having flights that were delayed. The data set that produces the initial highest R^2 value will be used for optimizing the model. This analysis will not include a leading coefficient, and will thus be left to the optimization stage.

Now that the data set has been chosen, the optimizations will take place. The first of the optimizations is to add a leading coefficient into the model to see if it affects the R^2 value. Next, interactions between variables will be modelled by creating a new column with the product of two other columns that could potentially interact with themselves. This will only be done on variables that are continuous, since discrete variables are made into dummy variables, and will not be affected by multiplication of another variable. Finally, each continuous column will be multiplied by itself and stored in a separate column to determine if non-linear regression has any significant increase in the accuracy of the model.

Polynomial regression however, is purely done for academic purposes to see the effect. A part of the analysis will be the creation of a residual plot to determine if the linear model is appropriate enough for this analysis. If the residual plot looks distributed around the horizontal randomly enough, then the linear regression will be deemed a good analysis technique.

IV. Results

1) With KNN Classification and using a k value of k = 655, the test set score was 78% which was the highest that we were able to obtain. The precision score for on-time flights was 78% while the recall was 100%. However, the precision and recall for delayed, diverted, and cancelled were all 0% as all flights were predicted on time.

   By lowering the k value significantly to k = 5, the test set score lowered slightly to 76%. The precision of on-time flights was 80% while recall was 93%. Delayed flights had a precision score of 35% with a recall of 16%. Precision and recall were both 0% for diverted flights while cancelled flights had 8% and 0% for precision and recall respectively.

   With the smallest possible k value of k = 1, the test set score dropped to 68%. The precision of on-time flights was still 80% while recall dropped down to 80%. Delayed flights had a precision of 27% and recall of 28%. Once again diverted flights had 0% for both scores. Lastly, cancelled flights had a score of 3% for both precision and recall. All of these results will be discussed in further detail in the subsection Key Findings.

2) The delay type classifier was unable to predict different delay types with confidence. Three different models were created to attempt to improve results: one model was built using multi-label classification, and the other two models used single-label classification based on the highest time delay for each delayed flight. The delay types with the best prediction results included carrier, late aircraft, and NAS delays. Security and weather delays reported below 10% for precision and recall throughout all tests. The most important, or influential, features in this analysis were determined to be the month, which may indicate seasonal influence on the model, as well as scheduled arrival and delay times, likely indicating daily air and ground flight congestion since most flights are planned for the middle of the day.

3) Although the analysis did not yield a model that could predict the length of the delay based on any scenario, as verified by the low R^2 value of .06, inspection into the coefficients did yield some interesting findings. The elements with the most positive affect on the delay time were a security delay type, flights travelling to Pensacola, Florida, and any flight within the Texas state. This makes sense because security delays may affect an individual passenger, but perhaps not the entire flight. Both airports in Pensacola, Florida and in Texas are also fairly close in distance to the George Bush airport, meaning delays could be minimized due to the distance. The elements with the most negative affect on the delay time were a weather delay type, flights travelling to Anchorage, Alaska, and any flight travelling to the Hawaiian state. Again, these make sense because a weather delay could ground a flight, continuously compounding a delay further and further. Hawaii and Alaska are as far away as can get in America, thereby compounding the delay due to distance. It's important to note that distance did indeed have a minor affect too, so any flight with a long travel distance will compound a delay. Finally, the length of the delay is not affected differently by different airlines. Each airline contributes the same amount to the length of a delay.

V. Discussion

A. Key Findings

1) With the use of KNN Classification, flight delays and cancellations could not be predicted with any degree of success based on the data analyzed. Using a k value of 655 (the square root of the size of the training set) the test score was determined to be 78%. While this may appear fairly accurate, by looking closer at the results

it becomes apparent that the reason for this score is due simply to the fact that each test point is predicted to be on-time as the vast majority of training points (78%) are on-time. By using such a large k value the model has become severely underfit and is ignoring any small but important patterns that may be present. As seen in the confusion matrix and classification report below, we can see that all the delayed, diverted, and cancelled test points were incorrectly classified as on-time. This led to a precision score of 78% and 100% recall score for on-time flights but 0% for both for all other classifications.

```
y-predicted:

On Time:  107121
Delayed:  0
Diverted:  0
Cancelled:  0

               On-time   Delayed  Diverted  Cancelled

On-time         83997       0        0          0
Delayed         21753       0        0          0
Diverted          297       0        0          0
Cancelled        1074       0        0          0

                          precision    recall  f1-score   support

On-Time:               0       0.78      1.00      0.88     83997
Delayed:               1       0.00      0.00      0.00     21753
Diverted:              2       0.00      0.00      0.00       297
Cancelled:             3       0.00      0.00      0.00      1074

             avg / total       0.61      0.78      0.69    107121
```

Fig. 2. Confusion Matrix and Classification Report for k = 655

Due to this underfitting, smaller k values were tested to determine if an optimal value could be found. Unfortunately, no test set score was found to be higher than 78%. As the k value was decreased, the model became less and less underfit. However, as shown below, even with a small k value of 5, few flights are predicted to be delayed, diverted or relative to the actual test set classifications (shown by the support score). With this k value, the precision of delayed flights was found to be 35% however, the recall score was only 16% as a result of not enough test points predicted as delayed. The precision and recall scores for diverted and cancelled flights were too low to be significant.

Finally, testing the data with the smallest possible k value of 1 further confirms just how random flight delays, diversions, and cancellations are. As one can see in the figure below, the ratio of predicted on-time, delayed, diverted, and cancelled flights is extremely close to the actual classifications of the test set (shown by the support score). However, the accuracy of these results are very low with only delays having precision and recall scores worth mentioning at 27% and 28% respectively. This does indicate that there may be some trends related to flight delays and this was analyzed in more detail for Research Question 3 below. At this k value, the ratio of false positives and false negatives is nearly equal for each classification and the test set

```
y-predicted:

On Time:  97294
Delayed:  9813
Diverted:  1
Cancelled:  13

               On-time   Delayed  Diverted  Cancelled

On-time:        77733      6253       1         10
Delayed:        18357      3394       0          2
Diverted:         267        30       0          0
Cancelled:        937       136       0          1

                          precision    recall  f1-score   support

On-Time:               0       0.80      0.93      0.86     83997
Delayed:               1       0.35      0.16      0.22     21753
Diverted:              2       0.00      0.00      0.00       297
Cancelled:             3       0.08      0.00      0.00      1074

             avg / total       0.70      0.76      0.72    107121
```

Fig. 3. Confusion Matrix and Classification Report for k = 5

score was 68%. Another trend that can be noted is that by decreasing the k value, the precision score of on-time flights stays fairly consistent while the recall score decreases due to more and more data points being incorrectly classified as other classifications.

```
y-predicted:

On Time:  83627
Delayed:  22126
Diverted:  276
Cancelled:  1092

               On-time   Delayed  Diverted  Cancelled

On-time:        67164     15792      206       835
Delayed:        15445      6019       68       221
Diverted:         227        67        1         2
Cancelled:        791       248        1        34

                          precision    recall  f1-score   support

On-Time:               0       0.80      0.80      0.80     83997
Delayed:               1       0.27      0.28      0.27     21753
Diverted:              2       0.00      0.00      0.00       297
Cancelled:             3       0.03      0.03      0.03      1074

             avg / total       0.69      0.68      0.68    107121
```

Fig. 4. Confusion Matrix and Classification Report for k = 1

2) The classification models built for predicting the type of aircraft delay did not provide a high enough accuracy and precision to confidently label the delay type.

One-vs-rest multi-label logistic regression was built to classify flights that may have more than one delay type. The model was run for each delay type independently, which allows a flight to be classified with more than one delay. Table I below shows the classification results for each delay type. As demonstrated in the table, the accuracy and precision average around 60% for viable classifications, such as carrier delay, NAS delay and late aircraft delay. Weather and security delays both show 0% precision and recall, which could be contributed to a lack of data points in the training and/or test set.

A single-label logistic regression model was built by classifying the delay type as described in section 2.C. This model resulted in an overall accuracy score of 42.9%. The maximum precision and recall scores were

TABLE I
MULTI-LABEL LOGISTIC REGRESSION CLASSIFICATION SUMMARY

Multi-label Logistic Regression				
Delay Type	Accuracy	Precision	Recall	F1 Score
Carrier	54%	54%	88%	67%
Late Aircraft	65%	51%	9%	15%
NAS	63%	64%	94%	76%
Security	98%	0%	0%	0%
Weather	91%	0%	0%	0%

for NAS delays, at 46% and 57%, respectively. The model also shows 0% precision and recall for security and weather delays, similar to the multi-label logistic regression model. Based on the results, this model does not provide a high level of confidence when predicting the type of delay. Table II below summarizes the results for the single-label logistic classifier.

TABLE II
SINGLE-LABEL LOGISTIC REGRESSION CLASSIFICATION SUMMARY

Accuracy	Delay Type	Precision	Recall	F1 Score
42.9%	Carrier	39%	38%	39%
	Late Aircraft	42%	37%	40%
	NAS	46%	57%	51%
	Security	0%	0%	0%
	Weather	0%	0%	0%
	Avg/Total	41%	43%	42%

The random forest classifier had revealed more promising and accurate results over the two prior logistic models. The classifier had an overall accuracy score of 43% and a maximum precision and recall score of 48% and 51%, respectively. Table III below shows the model results using a random shuffle of train and test data from delayed flights between the years 2003 to 2012.

TABLE III
RANDOM FOREST CLASSIFICATION SUMMARY

Accuracy	Delay Type	Precision	Recall	F1 Score
43.1%	Carrier	41%	42%	41%
	Late Aircraft	41%	40%	41%
	NAS	48%	51%	50%
	Security	3%	1%	2%
	Weather	6%	3%	4%
	Avg/Total	42%	43%	43%

Feature extraction was completed using the random forest classifier, and it was found that the most important features for building the model included: month, scheduled arrival time (crs_arr_time), and scheduled departure time (crs_dep_time). The importance of these features for classifying delay types could be season related as indicated by the month, and typical flight congestion in the air and ground as indicated by the scheduled arrival and departure time. The feature extraction also indicates that certain delay types are not generally specific to airlines, airports, or geographical regions. The figure below shows the top features and their corresponding influence on the model.

```
Variable: MONTH                Importance: 0.22
Variable: CRS_ARR_TIME         Importance: 0.22
Variable: CRS_DEP_TIME         Importance: 0.19
Variable: YEAR                 Importance: 0.07
Variable: DISTANCE             Importance: 0.03
Variable: DAY_OF_WEEK_1        Importance: 0.02
Variable: DAY_OF_WEEK_2        Importance: 0.02
Variable: DAY_OF_WEEK_3        Importance: 0.02
Variable: DAY_OF_WEEK_4        Importance: 0.02
Variable: DAY_OF_WEEK_5        Importance: 0.02
Variable: DAY_OF_WEEK_7        Importance: 0.02
Variable: UNIQUE_CARRIER_CO    Importance: 0.01
Variable: DAY_OF_WEEK_6        Importance: 0.01
```

Fig. 5. Feature Importance Using Random Forest Classifier Model

The random forest classifier was re-run using only the year, month, scheduled arrival, and scheduled departure features. The overall accuracy, average precision, recall and F1 score decreased when using the modified dataset; however, individual results improved for classifying the late aircraft, security and weather delays. Overall, the optimization of running the model using only the extracted features increases model efficiency while still providing output results that lie within reason to the original model. The results are demonstrated in table IV below.

TABLE IV
RANDOM FOREST CLASSIFICATION WITH FEATURE EXTRACTION SUMMARY

Accuracy	Delay Type	Precision	Recall	F1 Score
42.2%	Carrier	40%	39%	39%
	Late Aircraft	41%	42%	41%
	NAS	47%	50%	49%
	Security	8%	3%	4%
	Weather	8%	4%	5%
	Avg/Total	41%	42%	42%

Overall, this analysis did not prove to be an accurate measure to determine the type of delay an aircraft may be prone to; however, some meaningful results were found: the random forest classifier proved to be the best predictor for all flight delays; simplifying the model to use the most influential or important features gave similar results to the original model; the most important features in classifying delay types are month, scheduled arrival time, scheduled departure time, and year.

Improvements could be made to this model by diversifying the data-set to include information such as historical weather patterns, historical security information, and possibly flight maintenance information.

3) During the regression analysis, the R^2 value when using data that *didn't* know the type of the delay came to be .03. The R^2 when using data that *did* know the type of the delay came to be .09, which is approx. 200% higher. So this was the data set that was used going forward for optimizing.

   The R^2 value is still quite low, so the question that begs to be asked is whether or not this is an acceptable analysis or not. A residual plot was put together and can be seen below which shows a fairly random distribution of residuals among the horizontal. Since it's not as perfect as it could be, polynomial regression will be used to determine potential optimizations.

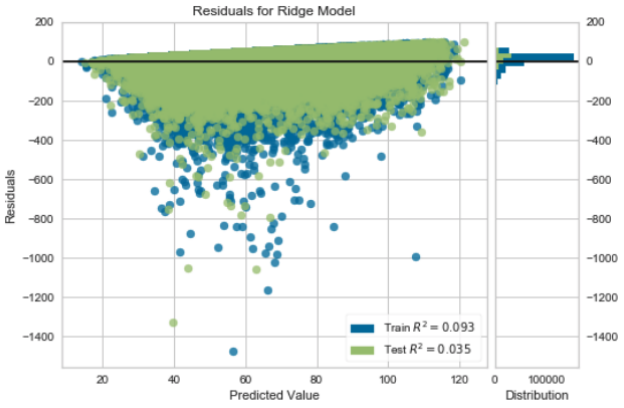

Fig. 6. Residual Plot

   The next optimization regarded including a leading coefficient in the model. The original model did not include one, and could very well explain the low R^2. This was done simply using scikit learn's regression module, which defaults to including a leading coefficient. Unfortunately there was no improvement.

   Next, both polynomial regression and variable interactions were added into the data set to see if they affected the model positively. This was simply done by including an additional column that carried the product of two other columns in itself. This could only be done on the continuous columns since the discrete columns only had either a 1 or a 0, which would not affect the model in any way. Unfortunately there was no improvement.

B. Limitations

Other methods in literature for predicting delays use delay propagation as a factor. This complex factor could not be included in the models created in this project. Adding this element in future studies could improve performance.

VI. Conclusion

Flight delays and cancellations are a serious economic concern for both the airlines and passengers. When planning a trip, the idea of knowing the likelihood of a flight being delayed/cancelled is very enticing so that one can plan accordingly. With access to millions of data points on domestic American flights from the year 1987 on, this project attempted to predict 3 aspects of flight delays.

First, a predictor was made and tested to determine whether a flight will be delayed/cancelled/diverted based on numerous variables such as destination state, destination airport, airline, and month using a KNN classifier. This classifier could not accurately predict the flight status.

Second, the type of delay could not be predicted with confidence using a random forest classifier or logistic regression. However it was discovered that month, scheduled arrival and departure time, and year had the most influence on the delay type.

Third, for flights that are delayed, regression was used to predict the total length of delay. It was discovered that for flights out of Houston, the elements with the most favourable effects on delay were the security delay type, Pensacola Florida airport, and Texas respectively. The least favourable effect was made by the weather delay type, Anchorage Alaska airport, and Hawaii state. Airline used had an equal effect on delay length.

The conclusions drawn from this study can help travellers, airports, and airlines improve the travel experience for everyone.

References

[1] TranStat, "Ostr bts transtats." [Online]. Available: https://www.transtats.bts.gov/HomeDrillChart.asp

[2] C. 30, "Types of delay," 06 2017. [Online]. Available: https://aspmhelp.faa.gov/index.php/Types_of_Delay

[3] "Airlines for america u.s. passenger carrier delay costs." [Online]. Available: http://airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/

[4] A. Guy, "Flight delays cost 32.9 billion, passengers foot half the bill," 10 2010. [Online]. Available: https://news.berkeley.edu/2010/10/18/flight_delays/

[5] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, *A Review on Flight Delay Prediction*, 03 2017.

[6] M. Hrastovec and F. Solina, "Machine learning model for aircraft performances," 05 2014.

[7] L. Zonglei, W. Jiandong, and Z. Guansheng, "A new method to alarm large scale of flights delay based on machine learning," in *2008 International Symposium on Knowledge Acquisition and Modeling*, Dec 2008, pp. 589–592.