AIAA Aviation
16-20 June 2014, Atlanta, GA
14th AIAA Aviation Technology, Integration, and Operations Conference

Predicting Ground Delay Program At An Airport Based On Meteorological Conditions

Avijit Mukherjee[1]
University Affiliated Research Center, Moffett Field, CA, 94035

Shon Grabbe[2] and Banavar Sridhar[3]
NASA Ames Research Center, Moffett Field, CA, 94035

In this paper, we present two supervised-learning models, logistic regression and decision tree, to predict occurrence of a ground delay program at an airport based on meteorological conditions and scheduled traffic demand. Predicting the occurrence of ground delay programs can help the Federal Aviation Administration traffic managers and airline dispatchers prepare mitigation strategies to reduce impact of adverse weather. The models are developed for two major U.S. airports: Newark Liberty and San Francisco International airports. The logistic regression model estimates the probability that a ground delay program will occur during a given hour. The decision tree model, on the other hand, classifies whether or not a ground delay program is likely during an hour based on the input variables. Results indicate both models perform significantly better than a purely random prediction of ground delay program occurrence at the two airports. The degree to which various input variables impact the probability of ground delay program vary between the two airports. While the enroute convective weather is a dominant factor causing ground delay programs at Newark Liberty Intl. airport, poor visibility and low cloud ceiling caused by marine stratus are major drivers of ground delay program occurrence at San Francisco Intl. airport.

---

[1] Research Scientist, Aviation Systems Division, MS 210-8, AIAA Member
[2] Research Scientist, Aviation Systems Division, MS210-10, Associate Fellow of AIAA
[3] Senior Scientist, Aviation Systems Division, MS210-10, Fellow of AIAA

I. Introduction

Adverse weather conditions such as poor visibility, low cloud ceiling, high winds, and convective weather cause capacity reduction at airports. A Ground Delay Program (GDP), which assigns pre-departure delays to aircraft inbound to the weather-impacted airport, is a control mechanism commonly used by air traffic managers to reduce arrival demand under such conditions. Predicting whether or not a ground delay program will be initiated at an airport, based on meteorological forecast and traffic demand, can alert traffic managers and airlines about potential congestion and necessitate strategies for mitigating these disruptions to air traffic. Machine learning methods [1, 2] can be applied to build such predictive models using historical data. The structure of the trained predictive models can provide insight into factors that influence the initiation of a ground delay program at a given airport. For instance, marine stratus at the morning hours is a major cause of ground delay programs at San Francisco International airport (SFO). High runway cross winds at Newark Liberty International (EWR) airport significantly influence initiation of ground delay program over there. Knowing the dominant factors causing ground delay programs at an airport can help focus development of technologies to improve the forecast accuracy of those factors and/or planning longer-term strategies to mitigate their impact.

Statistical models that measure the impact of meteorological conditions on performance metrics such as flight delays, cancellations, propagated delays, and passenger delays have been developed in the past. Ref. [3] provides an excellent review of literature on this topic. Unsupervised data modeling techniques such as *Principal Component Analysis*, and *Clustering* have been applied to classify days based on weather impact [4–6] and performance metrics [7]. While there are many applications of supervised learning methods such as logistic regression and decision trees, only a few studies have compared and contrasted the two [8–10]. Bloem and Bambos [11] applied *Inverse Reinforcement Learning* in predicting GDP occurrence. Inverse reinforcement learning is a relatively new technique within machine learning literature that attempts to model purposeful and strategic behavior of the decision makers. However, the results in Ref [11] indicate that GDP decision making is better modeled using traditional learning algorithm such as the ones used in this paper, compared to inverse reinforcement learning.The model presented in Ref [11] predicts whether or not a GDP will

occur during a time-interval, whereas the models presented in this paper determine the probabilities of occurrence of a GDP.

In this research, we develop models to predict the probability of occurrence of a ground delay program at an airport based on traffic demand and meteorological conditions. Two machine learning techniques, *Logistic Regression* and *Decision Tree*, are applied to develop these models. Both of these methods belong to the category of *supervised learning* in the literature of machine learning [2]. The models are applied to predict GDP occurrence at two major U.S. airports: EWR and SFO. Historical hourly observations of meteorological conditions such as visibility, cloud height, wind, convection, precipitation, etc., and arrival traffic demand based on flight schedules, are used to calibrate the models. The calibrated models are applied to predict GDPs on a test dataset. The logistic regression model estimates the probability of a GDP occurrence during an hour. The decision tree model, on the other hand, classifies the hour as a GDP or non-GDP. In this paper, we compare and contrast the performance of these two methods and discuss their applicability for predicting occurrence of GDPs.

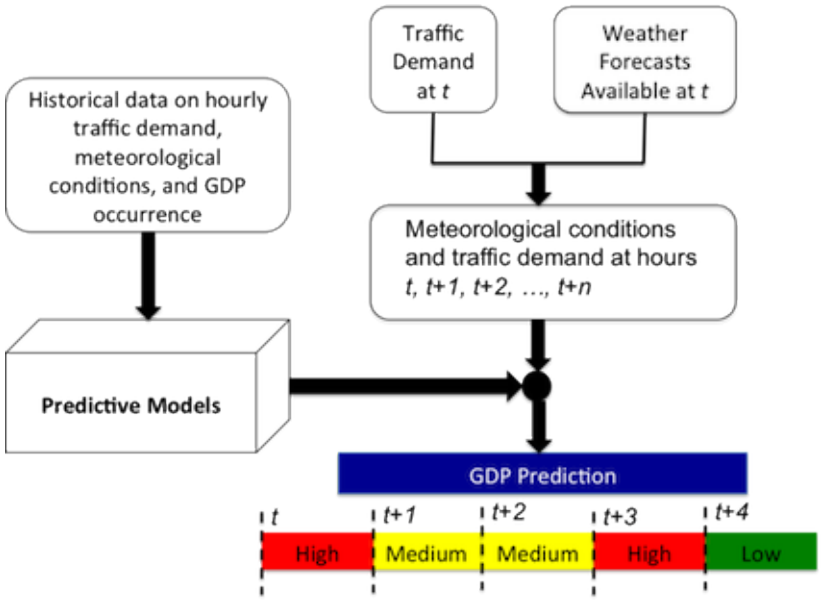
The outline of this paper is as follows. The main contributions and potential applications of this research are described in Section II. Section III describes the methodology. Section IV mentions the data and software used in this study. Section V presents the results, followed by conclusions in Section VI.

II. Motivation and Applications

The models developed in this research can be applied to predict the occurrence of ground delay programs at airports. Such predictive capabilities would help Federal Aviation Administration (FAA) traffic managers and airline dispatchers to prepare mitigation strategies for reducing traffic disruptions. The models are calibrated using historical data on meteorological conditions, traffic demand, and GDP occurrence. On a given day of operations the probability of a GDP at an hour in the future can be estimated using weather forecasts. In a simplistic setting, weather forecasts can be treated as deterministic. Uncertainty in forecasts can be accounted by calibrating the predictive models using both forecast and actual (i.e., nowcast) meteorological conditions. However, this is left

as a topic of future research. In this paper, we have chosen to calibrate the models using weather conditions that actually occurred. Figure 1 shows a prototype of a decision advisory system than can be built using the predictive models developed in this paper. At a given instant the most recent available weather forecast can be used to generate the hourly meteorological conditions. A GDP prediction can be accomplished by applying the models calibrated using historical data. Rather than providing the estimated probabilities of a GDP at a given hour, the decision support tool may display the chances of a GDP qualitatively, as shown in the figure. By setting appropriate thresholds on these probabilities one can classify the chances of a ground delay program under various conditions as high, medium, or low.

Fig. 1 Decision support system for predicting GDP



The parameter estimates of the explanatory variables in the logistic regression model measures the impact each variable has on the outcome (i.e., probability of a GDP). These values, along with their statistical significance levels, indicate the dominant meteorological factors causing GDPs at a given airport. In case of the decision tree, the path from the root node to each of the leaf node defines a set of conditions. The dominant class of a leaf node indicates whether or not the set of attributes, defined by the path from the root to the leaf node, are causing a GDP. The set of observations in each leaf node can be perceived as a cluster with similar attributes. While the observations in a cluster have similar meteorological and traffic conditions, TFM actions (i.e., GDP

in the current context) may not be the same for all of them. Investigation of clusters in which TFM actions to vary can be informative to the decision makers.

III. Methodology

In this section, we describe the application of the algorithms in predicting GDPs, and discuss how we evaluate their performances. For the sake of readability we present a succinct description of the applied machine learning methods, the key statistical parameters, and input data that is necessary to train the models.

A. Logistic Regression

The logistic regression model assumes that the probability, $p(\mathbf{X}_i)$, of GDP occurrence at an hour, i, is given by Eq (1). The components of the vector $\mathbf{X}_i$ are the explanatory variables such as meteorological conditions and traffic demand at that hour. The estimates of the coefficients βs, denoted by $\hat{\beta}$s, are obtained by maximizing the likelihood function [2] given by Eq (2). In the equation, n is the total number of observations and Y_i denotes the observed value of the binary response variable, i.e., GDP occurrence, at hour i.

$$p(\mathbf{X}_i) = \frac{e^{\beta^\top \mathbf{X}_i}}{1 + e^{\beta^\top \mathbf{X}_i}} \tag{1}$$

$$\mathcal{L} = \prod_{i=1}^{n} p(\mathbf{X}_i)^{Y_i} (1 - p(\mathbf{X}_i))^{(1-Y_i)} \tag{2}$$

Standard statistical software such as SAS, Matlab, and R provide modules and functions to estimate the parameters of a logistic regression model. While each application is unique in its own, there are a few key statistics that are used almost all the time. Along with the estimates of $\hat{\beta}$s, we obtain the Wald statistic given by $z = \frac{\hat{\beta}}{SE_{\hat{\beta}}}$, where $SE_{\hat{\beta}}$ denotes the standard error of $\hat{\beta}$. The square of the Wald statistic belongs to the χ^2 distribution with degree of freedom one. Therefore, the null hypothesis, $H_0 : \hat{\beta} = 0$, is rejected if z^2 is significantly greater than 0 (i.e., p-value is less than a threshold, which is usually set to 0.05). The log of likelihood ratio between full and reduced models indicates whether or not the full model (with larger number of explanatory variables) is a better fit; larger values generally indicate that the full model is a better fit. Another value of interest is the

odds-ratio for each of the explanatory variables. The odds-ratio of the k^{th} explanatory variable is given by $e^{\hat{\beta}_k}$, which is in fact the increase in $\frac{p(\mathbf{X}_i)}{(1-p(\mathbf{X}_i))}$ caused by unit increase in the variable X_{ik}. Odds-ratio is a key indicator of how much influence each explanatory variable has on the occurrence of GDP; a higher value indicates stronger influence. Other measures of model performance and statistical tests can be found in the literature.

B. Decision Tree

Before we describe how a decision tree classifies a dataset, we need to define some terminology. For a given group of observations, each of which belongs to a class, the impurity (also called as entropy) is given by Eq (3). In the present context, an observation contains a set of meteorological conditions and traffic demand, given by vector $\mathbf{X_i}$, during a given hour, i. The *class* of an observation is a binary variable that indicates whether or not there is GDP during that hour. In Eq (3), N denotes a node of the decision tree containing the group of observations, and ω_j is the fraction of observations that belongs to class j. Eq (3) can be generalized in cases where there are more than two classes. The decision tree algorithm splits the observations in node N into two subgroups (i.e., sub-nodes) by setting a threshold on a selected explanatory variable (also called as attribute) $X_{.k}$. Using a greedy heuristic, the attribute and its threshold are chosen so that the drop in impurity in the sub-nodes, given by Eq (4), is maximized. P_b in Eq (4) denotes the proportion of observations from node N that belong to sub-node N_b based on the split.

$$I(N) = -\sum_{j=0}^{1} \omega_j \log_2 \omega_j \tag{3}$$

$$\delta I(N) = I(N) - \sum_{b=1}^{2} P_b I(N_b) \tag{4}$$

The decision tree algorithm starts with the root node, which contains all observations of the input dataset, and recursively splits nodes until the impurity of leaf nodes is 0 (i.e., all observations in the leaf nodes belong to the same class). Forcing impurity at all leaf nodes equal to 0 can lead to over-fitting and an unnecessarily large number of nodes and branches in the decision tree. To prevent over-fitting, we apply a technique called *pruning*, in which, pairs of leaf nodes are joined until a performance criterion is reached. In this method, the input dataset is divided into two subsets:

calibration and validation. The decision tree is developed on the calibration set, and is grown to a full extent, i.e., reaching zero impurity at leaf nodes. It is then applied to the validation dataset and the misclassification rate, which is the fraction of observations misclassified by the tree, is computed. The misclassification rate usually reduces and then starts to increase as more leaf nodes are pruned. Leaf node pruning is done until the misclassification rate reaches a minimum. There are several variants of the decision tree algorithm [2]. They vary in defining the node impurity, node splitting criterion, pruning method, stopping criterion, and handling of missing data. In this paper, we use the *C4.5* algorithm, which was originally developed by J.R. Quinlan [12].

C. Comparing Logistic Regression and Decision Tree Models

The standard method to compare the performance of these two supervised classification algorithms is by computing their misclassification (i.e., error) rate on a new dataset, on which the models were not calibrated. The logistic regression provides a probability of occurrence of GDP based on the attributes of an observation. By setting a threshold on the probabilities, each observation can be classified as a GDP or non-GDP. For instance, if the threshold is set to 0.5, any observation whose estimated probability is less than this value will be classified as a non-GDP event, and those with estimated probability greater than 0.5 will be classified as GDP events. Then, by comparing classified values with the observed ones, we can calculate the error rate. By varying the probability threshold, one can obtain a distribution of the error rate. In a pruned decision tree, each leaf node is assigned a class, which is the class of some *proportion* of observations in that node. For instance, a leaf node can be classified to be "GDP" if 80% or more observations in the node belong to the class of GDP. The observations in the node that are non-GDP contribute to the error rate. As in the case of logistic regression, we can compute the misclassification rate by varying the threshold used to assign class to a leaf node. Whichever method generates a lower misclassification rate on the test dataset can be said to generate better predictions.

Another approach is to compare the Receiver Operating Curves (ROC) [1, 13] of the two models. The ROC curve plots the *true positive rate* vs. *false positive rate* of the classification. The area under the ROC curve indicates how well an algorithm classifies data compared to a "purely" random

classification. Larger areas indicate better predictive performance of the classification model. In this study, we apply the two classification algorithms on a test dataset and evaluate their performance.

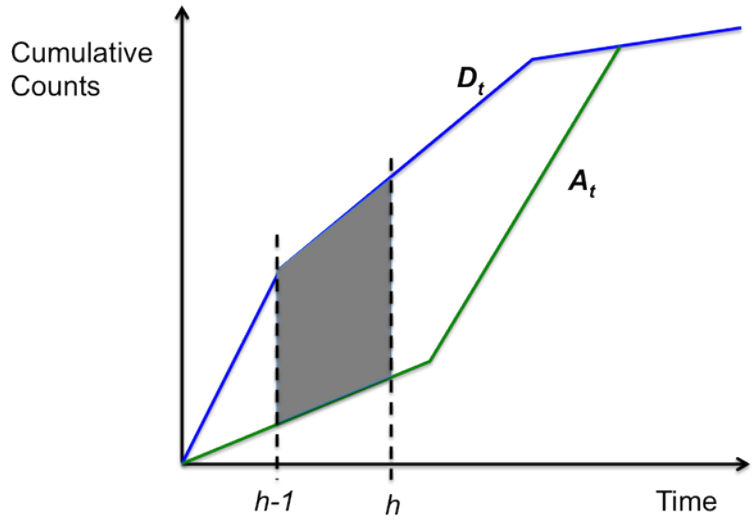
D. Input Variables and Model Calibration

So far we have mentioned the use of meteorological conditions as input to the predictive models. Here we describe them more specifically. As the input variables have different scales and variability, we use their standardized values as input to the predictive models. For each of the variable, V, we use their standardized values, given by $v = \frac{V-\mu_V}{\sigma_V}$, as input. μ_V and σ_V are the mean and standard deviation of V. Various meteorological conditions used in the models are as follows. Hourly observations of visibility, cloud height, and average wind speed are used. Magnitude of the head-wind and cross-wind components over the dominant runway configuration at an airport is computed from the observed wind speed and angle. Using convective weather and aircraft scheduled track data, we calculate the Weather Impacted Traffic Index (WITI)[4, 14] for various en-route Centers. The hourly WITI of the Center that encompasses the arrival airport are used as input. Along with the variables capturing meteorological conditions during a particular hour, the average of those values in the preceding three hours are also used as explanatory variables.

Flight arrival schedules at an airport are used to compute the hourly arrival traffic demand. Based on the observed visibility and cloud height at an hour, the airport arrival capacity is set to either the Visual Meteorological Conditions (VMC) or the Instrument Meteorological Conditions (IMC) capacity. Note that these values do not denote the actual operating capacity of the airport; rather, they are determined purely from the observed meteorological conditions. Using the hourly traffic demand and the assumed arrival capacity of the airport, we compute the *hypothetical* queuing delays [15] during various hours. The queuing delay computation is explained via Figure 2. The total queueing delay over a planning horizon is the area between the two curves: cumulative scheduled (D_t) and actual arrivals (A_t). The cumulative scheduled arrivals are simply derived from the hourly traffic demand. The cumulative actual arrivals at the end of hour h is given by: $A_h = min(D_h, A_{h-1} + C_h)$, where C_h is the airport arrival capacity at hour h, which is set based on the observed visibility and cloud ceiling, as described earlier. The area under the two curves between

$h-1$ and h is the queuing delay incurred during hour h.

Fig. 2 Queuing delay computation



Another derived variable, used as an input to the models, is the ratio of scheduled demand and the airport arrival capacity, C_h, which are determined based on the meteorological conditions (as explained above). We use this variable, denoted as ρ, instead of the scheduled traffic demand. The premise behind using ρ is that the chances of a GDP is higher when the same number of aircraft are scheduled arrive under worse meteorological conditions.

The input dataset consists of hourly observations (between 8AM - 11PM local time) of meteorological conditions, traffic demand, and an indicator of GDP being present, for calendar years 2011 and 2012. In total, there were 9552 observations. The predictive models are calibrated using 75% of the input dataset, and the remaining 25% is used for testing and comparing their performance. The sources of data and the software used in the experiments are described in the next section.

IV. Data and Software

Hourly observations of meteorological conditions at an airport were obtained from the Rapid Update Cycle (RUC) database. Traffic demand, based on flight schedules, were obtained from FAA's Aviation Systems Performance Metrics (ASPM) database. Data on GDP occurrence at an airport was obtained from FAA's National Traffic Management Log (NTML) database. The operating capacity of an airport under VMC and IMC were obtained from the FAA's Airport

Capacity Benchmark report [16]. The report also defines, for each airport, the meteorological conditions that classify an hour as VMC or IMC. SAS (version 9.2)[17] was used to develop the logistic regression model, and Weka [18] to develop the decision tree.

V. Results

In this section, we first present the calibration results of the predictive models, when applied to EWR. We then compare the performance of the two models in predicting GDP at EWR using a test dataset. As evident from the results, the logistic regression performs slightly better than the decision tree. Furthermore, the decision tree, even after pruning, is fairly large. We therefore chose to present only the logistic regression results for SFO. Using the results from the logistic regression model, we determine the influence of various input variables on the predicted probability of GDP at various airports.

A. Calibration Results of the Logistic Regression Model for EWR

The calibration result of the logistic regression model for EWR are presented in Figs. 3 and 4. The parameter estimates, $\hat{\beta}$s, indicate the how much influence each of the explanatory variable has on the probability of GDP at the airport. The $p - values$ indicate their statistical significance. The negative values of the $\hat{\beta}$ of visibility indicates that the probability of GDP reduces as visibility improves. While same explanation applies to cloud ceiling, its parameter estimate is not statistically significant. High cross-wind component over dominant runway configuration at EWR increases the chances of GDP. Based on the results presented in Fig. 3, the probability of GDP at a given hour is influenced most by the average ρ in the preceding three hours. If the average ρ in the preceding three hours increases by one standard deviation, the odds-ratio of GDP increases by 2.29. Among other variables, the average hourly WITI of New York Center (ZNY) during the preceding three hours has a significant effect on the chances of GDP at a given hour. This is expected as convective weather at ZNY warrants TFM initiatives at the three major airport in the New York Center.

Figure 4 compares the box plots of the probability of GDP, estimated by the logistic regression model, for hours where a GDP actually occurred versus hours where there was no GDP. The edges of the boxes are the quartiles, mean is shown using '+' sign, and the median is the middle bar in

Fig. 3 Logistic regression results for EWR

Variable	Parameter Estimates ($\hat{\beta}$)	Standard Error ($SE_{\hat{\beta}}$)	z^2	p-value	Odds ratio
Intercept	-0.89	0.03	791.61	<0.0001	
Visibility	-0.23	0.04	40.91	<0.0001	0.79
Cloud ceiling	-0.03	0.04	0.50	0.4784	1.01
Demand Capacity ratio (ρ)	0.18	0.05	15.21	<0.0001	1.21
Cross wind	0.09	0.04	6.83	0.0089	1.09
WITI of New York Center	0.29	0.06	23.96	<0.0001	1.35
Queuing delay	0.65	0.04	276.44	<0.0001	1.93
Avg. wind speed during past 3 hours	0.52	0.04	204.57	<0.0001	1.69
Avg. ρ during past 3 hours	0.83	0.05	281.37	<0.0001	2.29
Avg. WITI of New York Center during past 3 hours	0.67	0.07	104.62	<0.0001	1.96

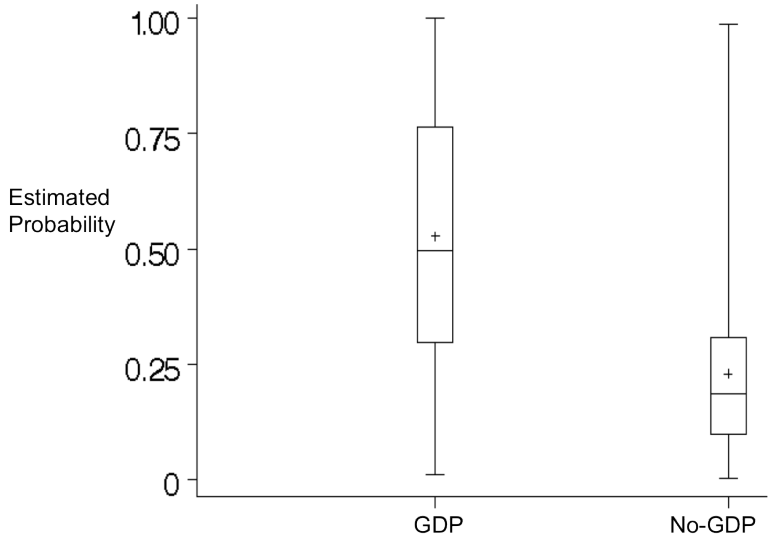
each box. As evident from the figure, the estimated probabilities in the cases where there was a GDP are, in general, higher than cases where there was no GDP. The average probability of GDP, estimated by the logistic regression model, in the hours where there was a GDP is approximately 0.5, whereas it is about 0.25 in cases where there was no GDP.

The null hypothesis of all βs being 0 simultaneously was rejected based on the Chi-square test of the log-likelihood ratio between full and reduced models. The area under the ROC curve was 0.83, which indicates that the estimated probabilities from the logistic regression model are significantly better than a "purely" random prediction, which would achieve an area 0.5.

B. Calibration Results of the Decision Tree for EWR

The unpruned decision tree has 491 nodes in total, among which, 246 were leaf nodes. After pruning, the size of the tree reduced considerably, with 93 nodes in total and 47 leaf nodes. A ten-fold cross validation method was applied to develop the pruned tree. A portion of the pruned decision tree for EWR is shown in Figure 5. In the figure, leaf nodes are colored in green. Although the re-scaled values of the input variables were used the build the tree, we describe the node splits using the actual values of the variables for better readability. Unlike logistic regression, the decision tree does not generate probability of occurrence of GDP. A set of conditions, defined by the thresholds of

Fig. 4 Box plot comparison of estimated probabilities



the variables that split various nodes in between the root node and a particular leaf node, classifies an observation as a GDP or non-GDP event. While the leaf nodes of an unpruned tree contain observations belonging to same class, a pruned tree misclassifies some observations. In Figure 5, the fraction of observations misclassified in a leaf node is shown in red color.

Figure 6 presents the true-positive and false-positive rates of predictions made by the decision tree for each observed class (i.e., GDP or non-GDP), along with their weighted averages across all observations. While the true-positive rate for non-GDP events is a fairly high value of 0.89, it is only about 55% for GDP events. This means that while the decision tree is correctly predicting non-GDP events almost 90% of the time, it is not able to do so well when predicting GDP events. The weighted values of true-positive and false-positive rates across all observations are 0.78 and 0.34 respectively. Area under the ROC curve based on the predictions from the decision tree on the calibration dataset is 0.80, which is slightly lower than that from the logistic regression model.

C. Performance Comparison Between The Two Models

The performance comparison of the two predictive models was conducted by applying them to the test dataset (described earlier). Figures 7 and 8 show the results. Based on the ROC curves presented in Figure 7, both models perform better than a purely random prediction. The area under ROC curves of the logistic regression model is 0.79, while that of the decision tree is 0.73. This

Fig. 5 Decision tree for EWR

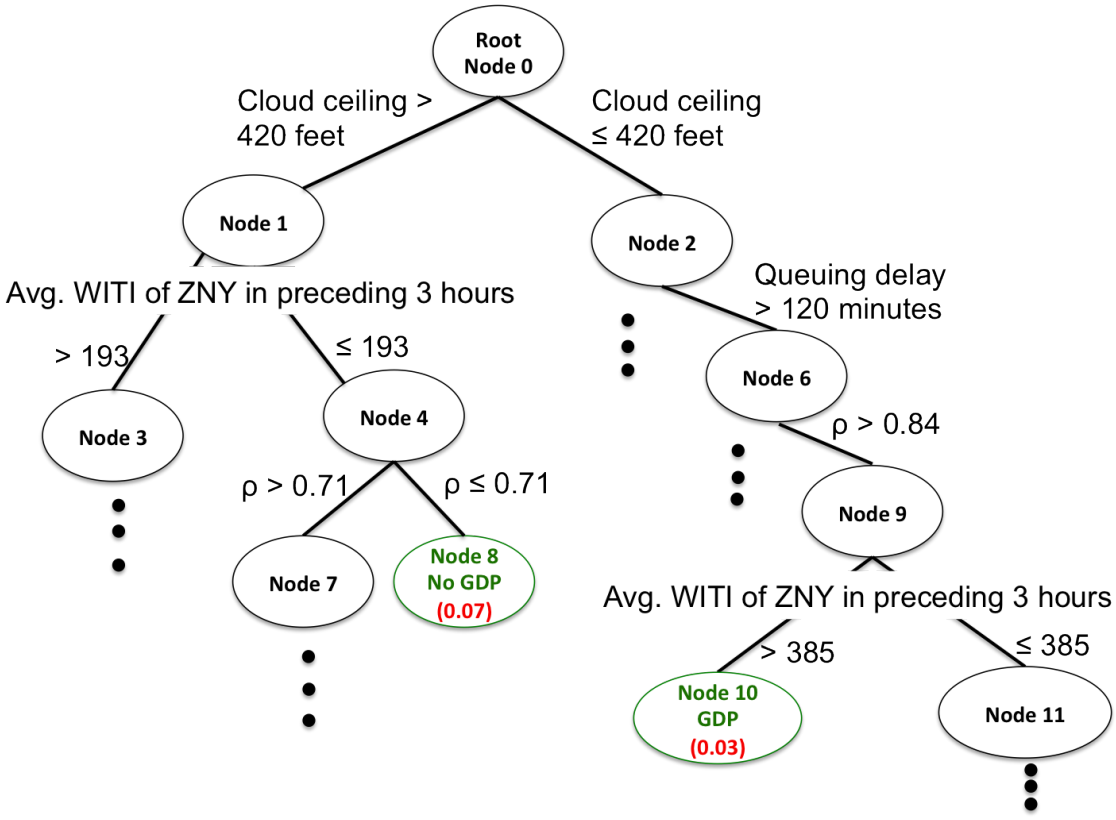


Fig. 6 Detection rates from decision tree applied to EWR calibration data

Observed Class	True-Positive Rate	False-Positive Rate
GDP	0.55	0.11
Non-GDP	0.89	0.45
Weighted average	0.78	0.34

indicates that the logistic regression performs better than the decision tree in predicting GDP on the test dataset. The error rate curves presented in Figure 8 confirms the same. The abscissa in the plot is the probability threshold for classifying an observation; as described in Section II.C. The threshold being close to zero are the cases where almost all observations are classified as GDP. The error rate of predictions from the logistic regression is generally lower than those from the decision tree. As the probability threshold increases from 0 to 1 the error rate for both models initially reduces and then

Fig. 7 ROC Curves for EWR Test Dataset

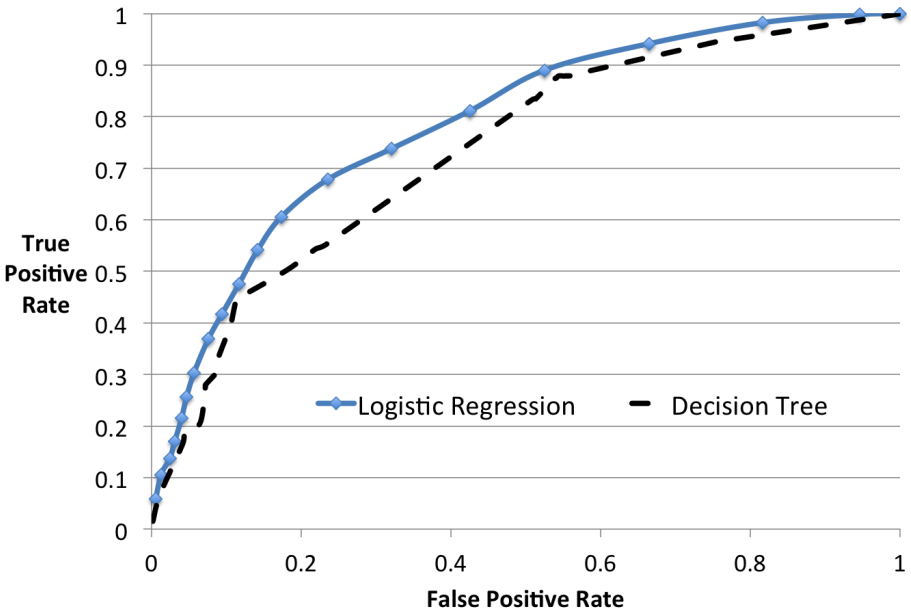
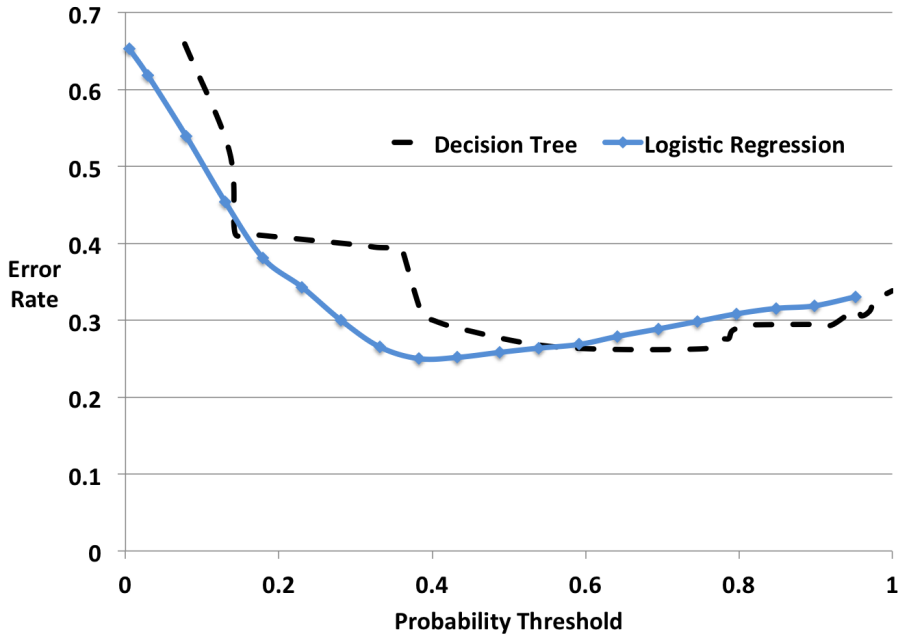


Fig. 8 Error Rate Curves



increases. For logistic regression, the error rate reaches a minimum around probability threshold 0.4, and then increases only slightly. For the decision tree, there is no clear minimum value of the error rate. Based on these results, we conclude that the logistic regression model is better suited in predicting GDPs at EWR than the decision tree.

D. Logistic Regression Model Results for SFO

This section discusses the performance of the logistic regression model applied to predict GDP events at SFO. Figure 9 presents the parameter estimates, their statistical significance, and the odds-ratio. As indicated by the negative values of parameter estimates, the probability of GDP at SFO reduces when visibility and cloud ceiling increases. Unlike in the case of EWR, the $\hat{\beta}$ of cloud ceiling is statistically significant with its $p-value$ less than 1%. The odds-ratios indicate that visibility and cloud ceiling influences the probability of GDP at SFO more strongly than they do at EWR. The result is fairly intuitive as the drop in visibility and cloud ceiling caused by marine stratus at SFO is a major factor that require GDPs at SFO. Another important difference between

Fig. 9 Logistic regression results for SFO

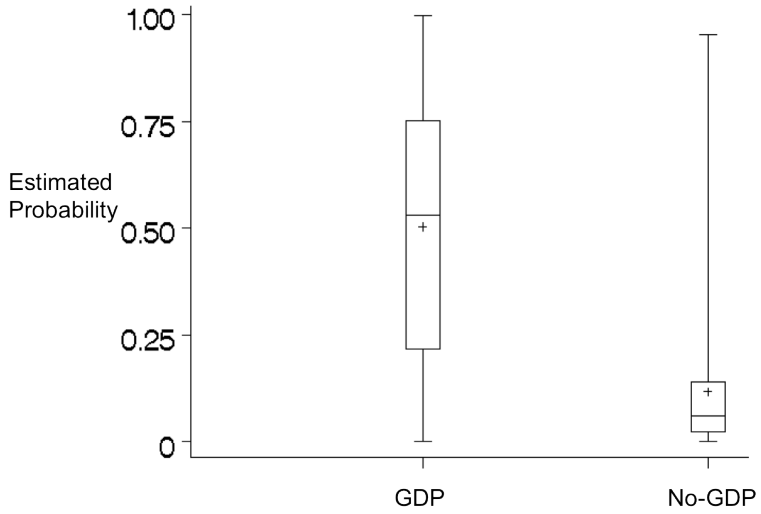
Variable	Parameter Estimates $(\hat{\beta})$	Standard Error $(SE_{\hat{\beta}})$	z^2	p-value	Odds ratio
Intercept	-2.29	0.05	1793.35	<0.0001	
Visibility	-1.42	0.08	302.65	<0.0001	0.24
Cloud ceiling	-0.41	0.05	67.60	<0.0001	0.66
Demand Capacity Ratio (ρ)	0.71	0.05	248.81	<0.0001	2.05
Cross wind	0.25	0.06	20.07	<0.0001	1.29
WITI of Oakland Center	0.07	0.04	3.24	0.0718	1.07
Queuing delay	0.17	0.04	18.68	<0.0001	1.19
Avg. wind speed during past 3 hours	0.14	0.05	8.18	0.0042	1.15
Avg. ρ during past 3 hours	0.61	0.05	191.69	<0.0001	1.85
Avg. WITI Oakland Center during past 3 hours	0.04	0.04	1.07	0.2994	1.04

SFO and EWR results is the impact of WITI of their respective Centers. While the hourly WITI of ZOA increases the chances of GDP at SFO, its statistical significance, indicated by the $p-value$, is less than that in the case of EWR. The same explanation applies to the average hourly WITI of the Oakland Center (ZOA) for preceding three hours. Among all variables, the ratio of demand and capacity (i.e., ρ) has the highest impact on the probability of GDP.

Figure 10 shows the box plot of the probability of GDP at SFO, classified by what actually occurred, for the test dataset. As evident from the figure, the estimated probabilities of GDP from the logistic regression model is generally higher for observations where GDP actually occurred

compared to the cases where GDP did not occur. By setting a threshold of 0.25 on the estimated probabilities, we can correctly classify GDP instances 84% of the time. If the threshold is set to 0.5, we can correctly classify GDP events almost 50% of the time. This implies that if the logistic regression predicts that a GDP will occur with 50% probability, the event actually occurs with almost the same probability. While this means that the false negative rate is about 0.5, the false positive rate is close to 0. The area under the ROC curve of GDP predictions on the test dataset is approximately 0.79, which indicates that the logistic regression model performs significantly better than a purely random prediction.

Fig. 10 Box plot of GDP probability of GDP at SFO for test dataset



VI. Conclusions

In this paper, we present two supervised-learning models, logistic regression and decision tree, to predict occurrence of a GDP at an airport based on traffic demand and meteorological conditions. The models are applied to predict GDP occurrence at two major U.S. airports: EWR and SFO. Historical hourly observations of meteorological conditions such as visibility, cloud height, wind, convection, precipitation, etc., and arrival traffic demand based on flight schedules are used to calibrate the models. The logistic regression model estimates the probability of a GDP occurrence during the hour. The decision tree model, on the other hand, classifies the hour as a GDP or non-GDP. We compare and contrast the performance of the two models and discuss their applicability in predicting the occurrence of GDPs.

Both models perform better than a purely random prediction of GDPs at the two airports. This conclusion is substantiated based on the area under their respective ROC curves. The logistic regression model, however, outperforms the decision tree model, but only by a slight margin. The decision tree model is complex and difficult to interpret. The degree to which various input variables impact the probability of GDP varies between the two airports. At both airports, input variables such as ρ, runway cross wind, and queuing delay, significantly impact the occurrence of GDPs. While ZNY WITI is an important factor impacting GDP at EWR, the ZOA WITI does not have such a strong influence on GDPs at SFO. While the enroute convective weather is a dominant factor causing GDPs at New York airports, poor visibility and low cloud ceiling caused by marine stratus are major drivers of GDP occurrence at SFO.

The models developed in this research can be applied to predict occurrence of GDPs at airports. Such predictive capabilities can help the Federal Aviation Administration (FAA) traffic managers and airline dispatchers to prepare mitigation strategies for reducing traffic disruptions. The models are calibrated using historical data on meteorological conditions and traffic demand. On a given day of operation the probability of a GDP at a specific hour in the future can be estimated using weather forecasts. In a simplistic setting, weather forecast can be treated as deterministic. Uncertainty in forecasts can be accounted by calibrating the predictive models using both forecast and actual (i.e., nowcast) meteorological conditions. Including weather forecasts as input variables is a direction of future research.

References

[1] Hand, D., Mannila, H., and Smyth, P., *Principles of data mining*, MIT Press, Massachusetts, 2001.

[2] Duda, R., Hart, P., and Stork, D., *Pattern Classification*, John Wiley and Sons Inc., New York, 2001.

[3] Sridhar, B., Grabbe, S., and Mukherjee, A., "Modeling and Optimization in Traffic Flow Management," *Proceedings of the IEEE*, Vol. 96, No. 12, 2008, pp. 2060-2080.

[4] Mukherjee, A., Grabbe, S., and Sridhar, B.,"Classification of Days Based on Weather-Impacted Traffic in the National Airspace System," *AIAA Aviation Conference*, Los Angeles, California, 2013.

[5] Asencio, M. A., "Clustering Approach for Analysis of Convective Weather Impacting the NAS," *12^{th} Integrated Communications, Navigation, and Surveillance Conference*, Herndon, Virginia, 2012.

[6] Liu, P-c. B., Hansen, M., and Mukherjee, A.,"Scenario-Based Air Traffic Flow Management: From Theory to Practice," *Transportation Research - Part B*, Vol.42, 2008, pp. 685-702.

[7] Hoffman, R., Krozel, J., Penny, S., Roy, A., and Roth, K., "A Cluster Analysis to Classify Days in the National Airspace System," *AIAA Guidance, Navigation, and Control Conference*, Austin, Texas, 2003.

[8] Bloem, M., Hattaway, D., and Bambos, N.,"Evaluation of Algorithms for a Miles-in-Trail Decision Support Tool," 5^{th} *International Conference on Research in Air Transportation*, Berkeley, California, 2014.

[9] Long, W., et al., "A Comparison of Logistic Regression and Decision-Tree Induction in a Medical Domain," *Computers and Biomedical Research*, Vol. 26, 1993, pp. 74-97.

[10] Rudolpher, S. M., Paliouras, G., and Peers, I.S., "A Comparison of Logistic Regression and Decision-Tree Induction in the Diagnosis of Carpal Tunnel Syndrome," *Computers and Biomedical Research*, Vol. 32, No. 5, 1999, pp. 391-414.

[11] Bloem, M., Bambos, N., "Ground Delay Program Analytics and Behavioral Cloning with Inverse Reinforcement Learning," *AIAA Aviation Conference*, Altanta, 2014.

[12] Quinlan, J.R., "Induction of Decision Trees," *Machine Learning*, Vol. 1, 1986, pp. 81-106.

[13] Hanczar, B. et al., "Small Sample Precision of ROC-Related Estimates," *Bioinformatics*, Vol. 26, No. 6, 2010, pp. 822-830.

[14] Sridhar, B. and Chen, N., "Short-Term National Airspace System Delay Prediction Using Weather Impacted Traffic Index,"*Journal of Guidance, Control, and Dynamics*, Vol. 32, No. 2, 2009, pp. 661-675.

[15] Daganzo, C., *Fundamentals of Transportation and Traffic Operations*, Elsevier-Pergamon, Oxford, U.K., 1997.

[16] Federal Aviation Administration., *Airport Capacity Benchmark Report*, U.S. Department of Transportation, Washington D.C., 2004.

[17] SAS Institute, *SAS/STAT (R) Version 9.22 User's Guide*, Version 9.2, North Carolina, 2010. URL: http://support.sas.com/

[18] The University of Waikato, *Waikato Environment for Knowledge Analysis (Weka)*, Version 3.6.10, New Zealand, 2013. URL: http://www.cs.waikato.ac.nz/ml/weka/