Predict the delay in your airline before they do!!

Hari Hara Sudhan, Rajesh Inbasekaran
Kavi Associates, Barrington, IL

Abstract

This paper demonstrates the application of predictive modeling techniques to predict the departure delay in several domestic flights across USA. Delay in domestic flights has been a common phenomenon in USA and it would be definitely useful if a predictive methodology was employed. The data set for this purpose was prepared by gathering the past 2 year's data from flight stats website. The weather details of these airports were also collected to understand if the weather details can be used for the prediction. By using modeling techniques such as Multiple Regression, Neural Networks etc., the delay in airlines can be predicted by knowing the airline carrier, origin and destination airport.

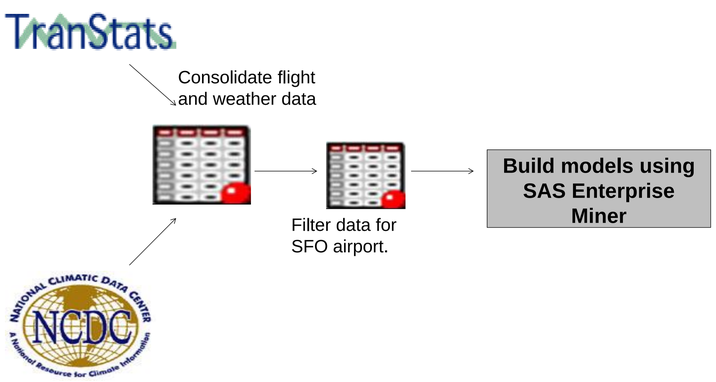
Introduction

This research paper analyses the transport statistics of all flights flying within USA for the period August 2010 to August 2012. The data is collected only for non-stop flights and the analysis does not include diverted and cancelled flights. The delay in departure is calculated by the difference between actual and estimated departure time. This project focuses only on the delay in departure and not on arrival.

A major factor for an airline delay is the weather. The weather details for these airports were also collected from National climate data center (NCDC). The weather details were available only for 180 airports of the 315 airports for which the flight stats were available. Based on these data the airline delay can be analyzed for each airport.

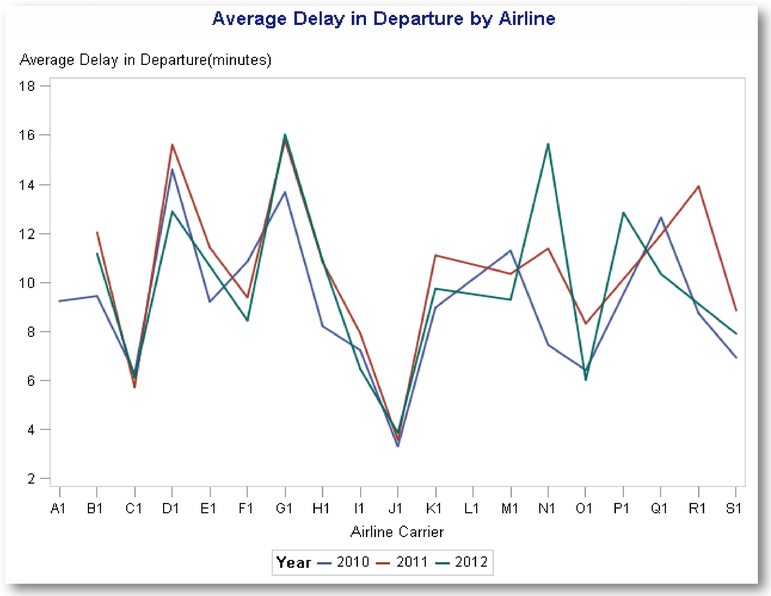
Data Preparation

The data were collected from transtats.gov which had the flight statistics for each month starting August 2010 to August 2012 in xls format. SAS Enterprise Guide was used in integrating the flight details for all months into a single SAS Dataset. Further details with respect to Airport and airlines were collected such as Airport's latitude and longitude information, City, State and Airport Id used by FAA and these were converted to SAS Datasets.

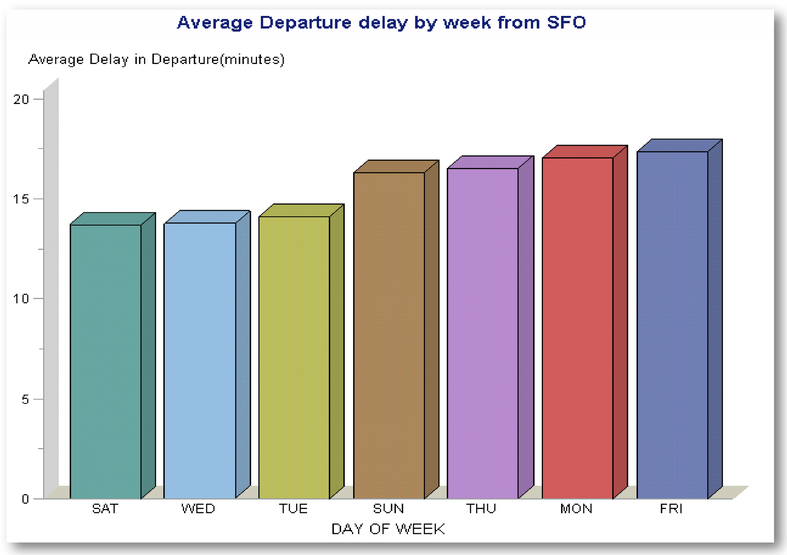


Descriptive Statistics

The data was analyzed using SAS Enterprise guide. Various airlines were analyzed for their delay(minutes) in departure over the three years. Following graph shows the average delay(in minutes) for all airlines across all airports. Airline F1 have reduced their delay in departure in 2012 compared to 2010 and 2011. Departure delay for airline N1, P1 has increased in 2012 as compared to 2010 and 2011.



Following graph shows the average delay in departure in a week for the San Francisco International airport (SFO). By the result shown below, we can say that flights leaving on Tuesday, Wednesday and Saturday has a low delay in departure compared to those leaving on other days of the week.

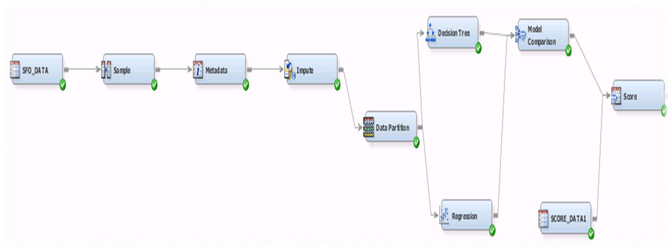


The above graph was performed only for SFO airport. Such analysis can be extended to other airports and analysis can also be done based on the day of the month and the month of the year. Such analysis gives more insight on which month of the year or day of the month there occurs more flight traffic and thereby causing airlines to delay their flight departure.

Data Mining:

SAS Enterprise Miner 12.1 was used for building the models. The data was filtered for San Francisco International airport (SFO) details. The filtered data set had 300,000 records containing details of all flights departing from SFO. In the airline industry a flight is said to be on-time if it is within 15 minutes of its estimated departure time. The target variable was prepared such that it had either 0 or 1, 0 specifies flight was on time or less than 15 minutes of its estimated departure time and 1 being there was a delay which is 15 minutes more than estimated departure time.

A 70-30 partition was used to split the data to training and validation datasets.



Several models like Decision tree, Neural network, Regression were tried and based on the Misclassification rate, Logistic regression was chosen as the best model. The validation MISC rate for the model was 21.3%.

Step	Effect Entered	Label	Pr > ChiSq
1	DepTimeBlk	Departure Time frame	<.0001
2	Precip_water_equiv	Water precipitation	<.0001
3	Airline	Airline carrier	<.0001
4	IMP_average_dew_point	Average dew point	<.0001
5	maximum_temp	Maximum temp.	<.0001
6	minimum_temp	Minimum temp	<.0001
7	dest_Airport_Name	Destination airport	<.0001

The above table shows the important variables that were selected by the regression model. Forward selection method was chosen for the regression method to select the variables. The fit statistics for the model is shown below.

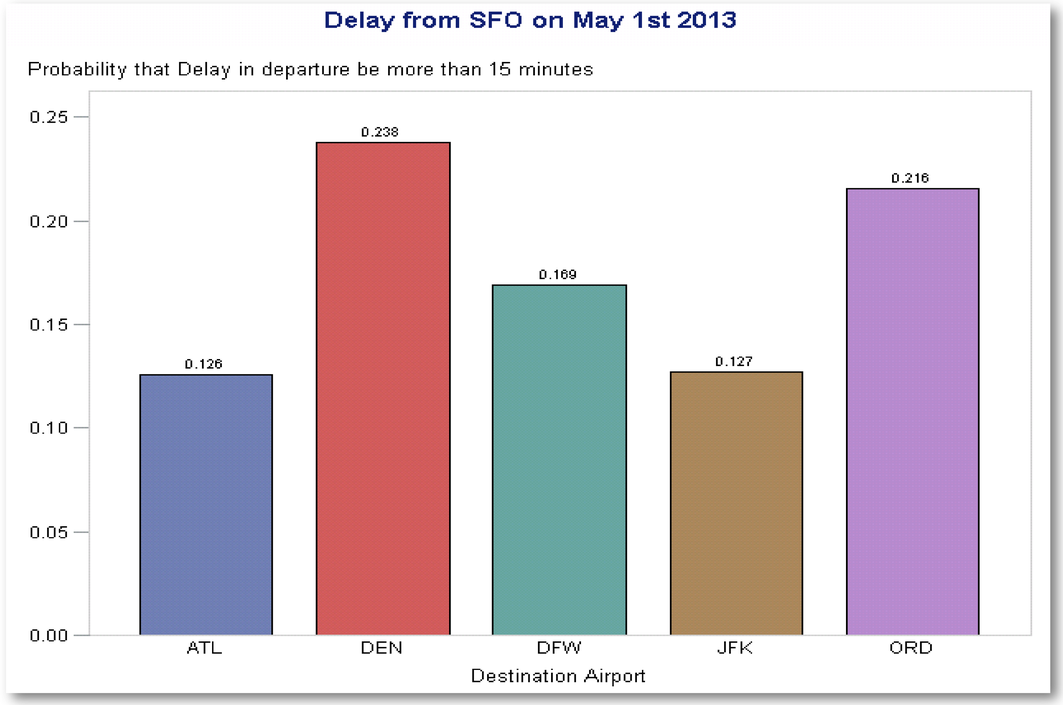
Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	25877.29	.
ASE	Average Squared Error	0.153674	0.155197
AVERR	Average Error Function	0.477393	0.481334
DFE	Degrees of Freedom for Error	26753	.
DFM	Model Degrees of Freedom	113	.
DFT	Total Degrees of Freedom	26866	.
DIV	Divisor for ASE	53732	23032
ERR	Error Function	25651.29	11086.09
FPE	Final Prediction Error	0.154972	.
MAX	Maximum Absolute Error	0.989404	0.988021
MSE	Mean Square Error	0.154323	0.155197
NOBS	Sum of Frequencies	26866	11516
_NW_	Number of Estimate Weights	113	.
RASE	Root Average Sum of Squares	0.392013	0.39395
RFPE	Root Final Prediction Error	0.393665	.
RMSE	Root Mean Squared Error	0.39284	0.39395
SBC	Schwarz's Bayesian Criterion	26803.73	.
SSE	Sum of Squared Errors	8257.203	3574.491
SUMW	Sum of Case Weights Times Freq	53732	23032
MISC	Misclassification Rate	0.211606	0.213529

4 ft. x 2 ft. poster

TEMPLATE

Predicting the Delay in Departure on May 1st 2013 from SFO:

The regression model was used in predicting the delay in departure for all airlines departing from SFO on May 1st 2013. Based on the predictor variables and the weather information the chances of the delay to selected destinations can be predicted irrespective of the airline.



Conclusion:

Delays in flight departure can be subjected to various reasons. A lot of factors goes into predicting a delay in a flight departure. This paper focuses on one such factor, weather. Several factors can be identified and data related to those can be collected and can be used to build various models to better predict the delay in a flight across all airports in USA. A wide variety and a rich collection of data would definitely be useful in building a better model to predict the delay.

References

Airline On Time Performance *Arzu Yesilova, Brad Peters,Irem Ataibis,Kanokporn Laochunsuwan,Sergio Iovanovich.*

Acknowledgments

The authors would like to thank National Climate Data center and www.transtats.bts.gov for providing us with the data to carry out the project.

Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Hari Hara Sudhan
Company: Kavi Associates
E-mail: hari.duraidhayalu@kaviglobal.com

Name: Rajesh Inbasekaran
Company: Kavi Associates
E-mail: rajesh@kaviglobal.com