Comparison of Prediction Power of Models used to Predict Flight Delays at Jomo Kenyatta International Airport

Abstract

Delays in flights have negative socio-economics effects on passengers, airlines and airports, resulting to huge economic loses. Therefore, their prediction is crucial during the decision-making process for all players of aviation industry for proper management. The development of accurate prediction models for flight delays depend on the complexity of air transport system and airport infrastructure, hence may be country specific. However, there exists no prediction models tailored to Kenyan aviation industry. Hence there is need to develop prediction models amenable to Kenya aviation conditions. The objective of this study was to compare the prediction power of the developed models. Secondary data from JKIA for the period from March 2017 to March 2018 was used. The data collected included the day of the flight (Monday to Sunday), the month (January to December), the airline, the flight class (domestic or international), season (summer or winter), capacity of the aircraft, flight ID (tail number) and whether the flight had flown at night or during the day. The analysis of the data was done using R- software. Three models, Logistic model, Support Vector Machine model and Random Forest model, were fitted. The strength and utility of the models was determined using bias-variance learning curves. The study revealed that the models predicted delays with different accuracies. The Random Forest model had a prediction accuracy of 68.99% while the Support Vector Machine model (SVM) had an accuracy of 68.62% and the Logistic Regression model had an accuracy of 66.18%. The Random Forest model outperformed the SVM and Logistic Regression with accuracies of 0.37% and 2.71% respectively. The SVM and Random Forest do not assume probability distribution of the response under investigation, probably indicating why they performed better than the logistic regression. The study recommends application of Random Forest model to predict flight delays at JKIA.

Key Words: Flight Delays; Prediction Model; Support Vector Machine; Logistic Regression Model; Random Forest Model

Introduction

Statistical modelling is a mathematical way of making approximations from input data. These approximations are then used to make predictions (Lunt, 2013). Statistical models help in predicting the future probabilistic behaviour of a system based on past statistical data (Waljee *et al.,* 2014; Geisser, 2016). Predictive modelling has been used in many fields, for example in crime cases (Finlay, 2014); to detect the likeliness of an email being spam (Sheskin, 2011) and flight delays (Kalliguddi & Leboulluec, 2017).

In evaluation of how different models perform in modelling of flight delays, regression models have been found efficient in predicting flight delays since they highlighted the various causes of flight delays (Sternberg *et al.*, 2017). However, they could not categorize complex data. Econometric models have been used to model scheduled flight cancellation and to show how delays from one airport were propagated to other destinations (Hao *et al.,* 2014). These models did not provide a complete vindication since they ignored variables that were difficult to quantify. When subjected to social-economic situations, the models showed discriminative and subjective results (Hao *et al.,* 2014). Among the models used, random forest has been found to have superior performance (Rebollo & Balakrishnan, 2014). Prediction accuracy may vary due to factors such as time of forecast and airline dynamics. A developed multiple regression model has shown that distance, day and scheduled departure are key factors in predicting flight delay (Burgauer & Peters, 2000). However, though the model gives flagged out the significant factors, its prediction accuracy was poor. Moreover, the model is limited to only one flight route (Burgauer & Peters, 2000).

Comparison of other models, such as the K-means clustering Algorithms and Fourier fit model, have shown that Fourier fit model could predict flight delays with a high precision (Qin & Yu, 2014). However, the two models were found to be suitable a single airport, but not prediction applied to multiple airports. Probability models such as the normal distribution and the Poisson distribution have been used to model flight departure and arrival delays (Mueller & Chatterji, 2002). However, the prediction accuracy varied depending on variables such as time duration and the number of airports considered. Normal distribution was observed to model flight departure delays better while arrival delays were modelled better by the Poisson distribution (Mueller & Chatterji, 2002). However, these models are parametric and assume that the response takes a particular functional form. If this form is not met by the training data set, the resulting model will not fit the data well and the estimates from this model will be poor.

Logistic regression model has been used to model flight on-time performance (Arjun *et al*., 2013). The model showed good performance with the training data set and the testing data set. The variance of the model was also low. However, its parametric nature can be a weakness if the training data set will not meet the assumed functional form. Neural networks performed better than logistic regression model in prediction of death in patients with suspected sepsis in an emergency room (James *et al*., 2013). This was attributed to the neural networks having few features to be verified before model construction and its ability to fit non-linear relationship between dependent and independent variables. Support Vector Machine (SVM) model was fitted and it was observed to fit all the training data set correctly (Arjun *et al*., 2013). In prediction of auto-ignition temperatures of organic compounds, SVM perfomed better than multiple linear regression and back propagation neural network (Pan *et al.,* 2008). Random forests have been used to model delay innovation (Rebollo & Balakrishnan, 2014). Results from this study showed that more decision trees were better but up to a certain critical value. Prediction of new vehicle prediction approach in computational toxicology led to results with random forest performing better than decision tree (Mistry *et al*., 2016).

Random forests and SVM are classified under machine learning. Under machine learning, the training data is divided into several samples (Tripathi & Naganna, 2015). At each sample, a model is fitted and tested against the testing data set. The sample that yields the best model is obtained from a plot of the train errors and the test errors against the sample size (Sui *et al*., 2003). Their overall advantage of the SVM and the random forest is their non-parametric nature in that they do not assume a particular functional form of the response under investigation. This makes them very flexible since they fit a wider range of shapes of the response (Cristianini & Shawe, 2000). Modelling studies on flight delays are not available for Kenya aviation industry. The aim of this study is to compare the prediction power of models that have been used to predict flight delays at Jomo Kenyatta International Airport.

Methodology

Secondary data that was obtained from Kenya Airports Authority on flights at Jomo Kenyatta International Airport. The data was for the year 2017/2018 where the year started on March 2017 and ended on March 2018. The variables used included; the day of the flight (that is, Monday to Sunday), the month (that is, January to December), the airline, the flight class (that is, domestic or international), season (that is, summer (March to October) or winter (October to March), capacity of the aircraft, flight ID (tail number) and whether the flight had flown at night or during the day. The data was analysed using R-core statistical software. The time difference between the scheduled time and the actual time for flights was calculated. A time difference of more than 15 minutes was classified as a delay and it was given a value 1 and a time difference of less than 15 minutes was classified a non-delay and given the value 0. The three models, logistic regression model, SVM model and Random Forest, were fitted by machine learning. The entire data set was divided into a training data set of 15000 flights and a testing data set of 5000 flights. In fitting the models, different random samples were created from the training data by the programmed laptop used. For each sample, a model was fitted and tested using the testing data. The entire process of developing the models is summarised in Figure 1.

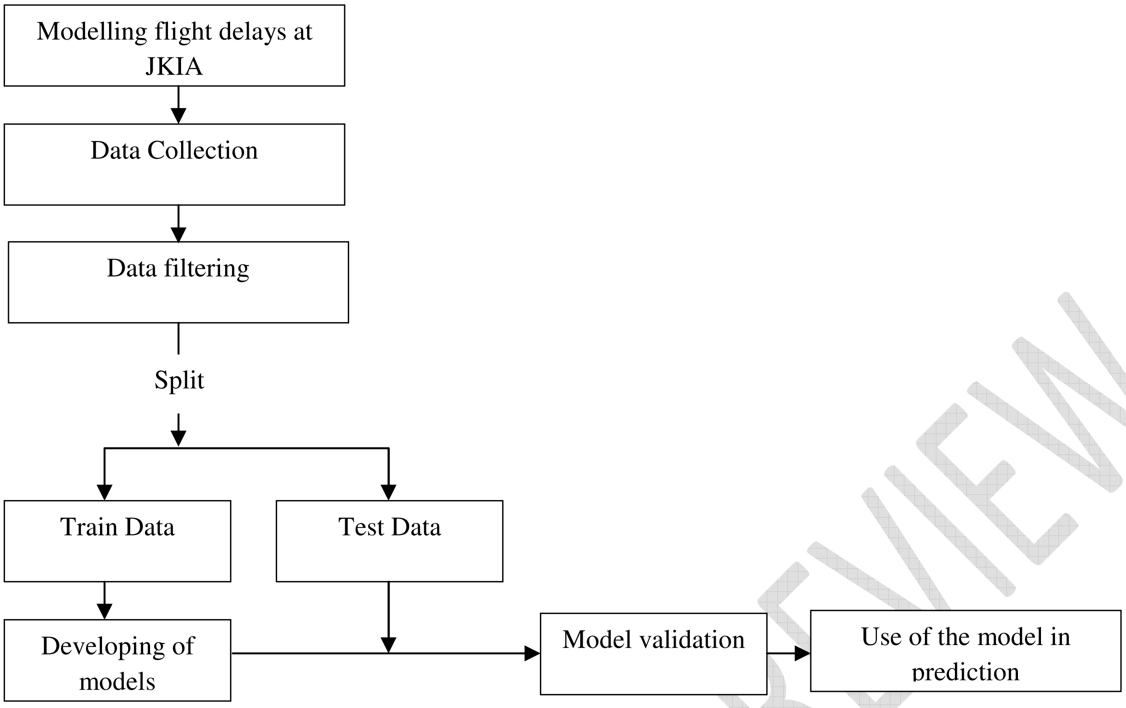

Figure 1: Schematic Diagram of the Modelling Process

Models Comparison

One of the techniques that was used to compare the fitted models was the bias-variance curves. Bias-Variance curves showed how each model performed with both the train data and test data (Meek *et al*., 2002). Bias measured how the average accuracy of an algorithm changes as the input data changes. Variance measured how sensitive the algorithm is to the chosen input data. Learning curves were plotted using errors from the training data set and the testing data set on the same axes. Learning curves were used to evaluate the performance of a model with both the train data and the test data (Meek *et al*., 2002)

Another technique that was used in models comparison was use of accuracy, precision, recall and $F1$ score. These statistical terms can be defined according to Olson *et al.* (2008).

$$\text{Accuracy} = \frac{\text{True Positive+True Negative}}{\text{Total Population}}$$

$$\text{Precision}=\frac{\text{True Positive}}{\text{True positive+false positive}} = \frac{\text{True Positive}}{\text{Total Predicted Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} = \frac{\text{True positive}}{\text{Actual Positive}}$$

$$F1 \text{ score} = 2*\frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}}$$

Accuracy cannot on its own be completely used when determining the best model (*David,* 2011). Precision indicates how many of the positively predicted observations are actually positive. Precision is thus very useful when determining the best model when the cost of false positive is high. Recall indicates how many of the actual positives are true positives. Recall is thus important when choosing the best model in a case where the cost of false negative is high. The $F1$ score strikes a balance between precision and recall. Accuracy can be largely contributed by a large number of true negatives which in most business situations are not very important. In such cases, $F1$ score becomes a better way of determining the best model (Fawcett, 2006).

Results and Discussion

Models comparison using Bias-Variance curves

i) Fitted Logistic Regression Model

The best logistic regression model was obtained after a sample size of 5000 observations (Figure 2). In a study by Arjun *et al*. (2013), a bias-variance learning curve showed convergence after 2500 observations. This implies that for different studies, the best model can be obtained at different sample sizes.

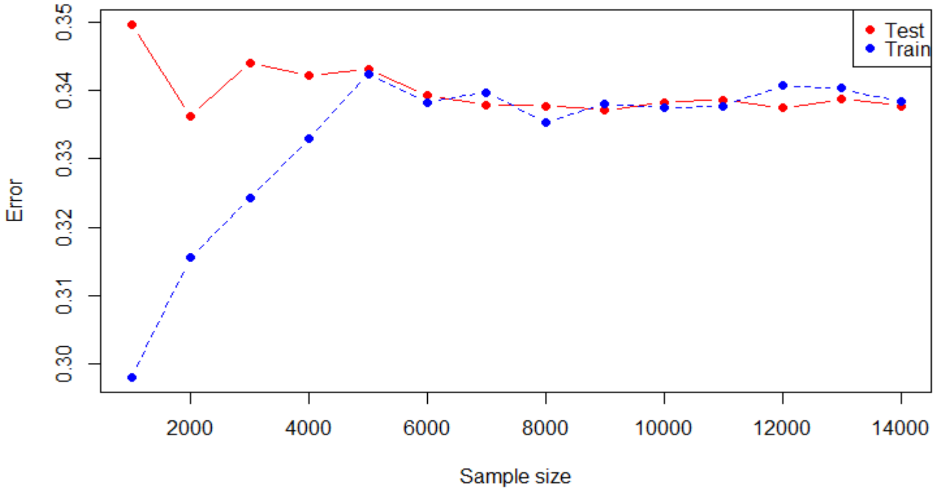

Figure 2: Bias-Variance Curve for Logistic Model

ii) Fitted Support Vector Machine Model

The support vector machine model obtained in this study was fitted using the e1071 package implemented in R software. A good SVM model was obtained for a sample size of 5000 (Figure 3). However, increasing the sample size beyond 5000 could not improve the model. A good logistic model was also obtained at a sample size of 5000. This indicates similar performance in the logistic model and the Support Vector Machine model in predicting flight delays. The accuracy value implied that the SVM model predicted whether a flight could be delayed or not with an accuracy of 68.28%

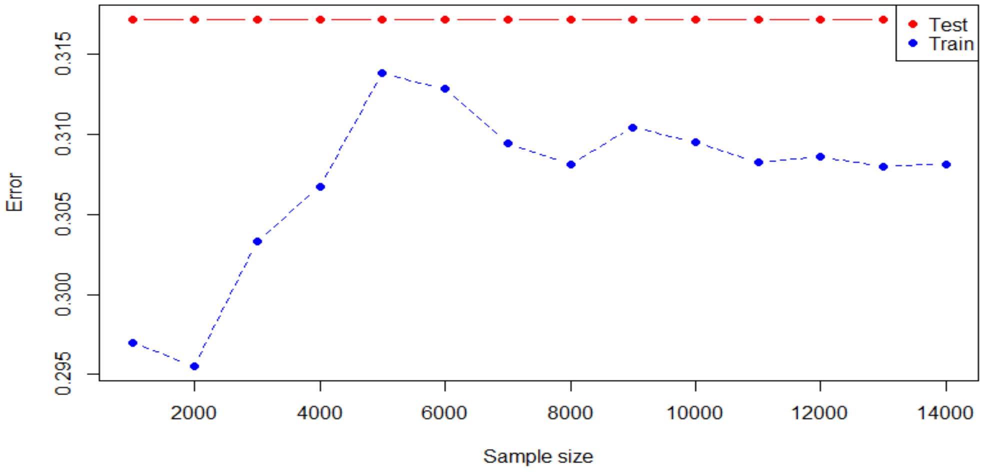

Figure 3: Bias-Variance Curve for Support Vector Machine Model

Fitted Random Forest Model

This model was implemented using party package in R software. The fitted random forest model had an accuracy of 66.99% in predicting whether a flight was delayed or not delayed. A good random forest model was obtained for a sample size of about 3000 flights (Figure 4). This implies that the random forest model achieved a better model with a smaller sample as compared to the logistic regression model and the support vector machine model. Analysis of financial credit risk using machine learning, random forest achieved a better model with a smaller sample compared to support vector machine and logistic regression model (Chow, 2018). This indicates that, the random forest model requires a smaller sample to fit a good model when compared to Support Vector Machine and Logistic Regression models.

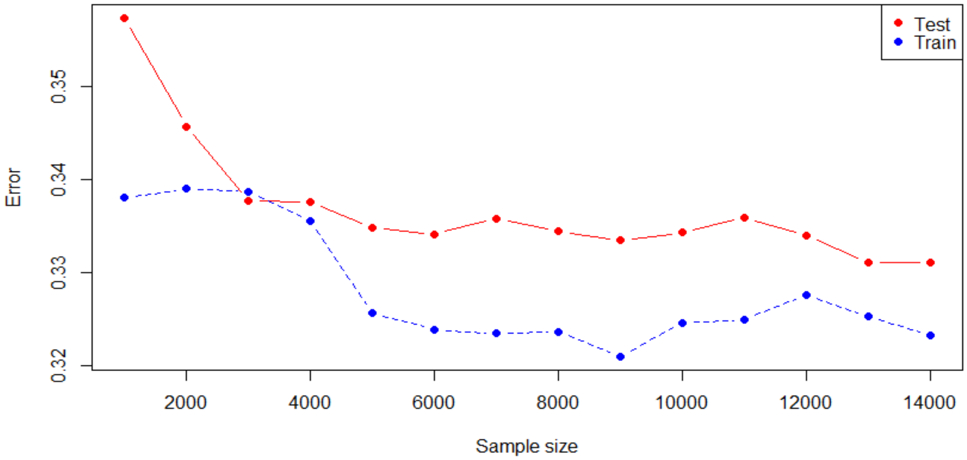

Figure 4: Bias-Variance Curve for Random Forest

Comparing the fitted Models using Accuracy Tables

The fitted models were compared using their values of accuracy and $F1$ score (Table 1). The random forest had an accuracy of 0.6899 and $F1$ score of 0.7793. The support vector machine had an accuracy of 0.6862 and $F1$ score of 0.7764. The logistic regression model had an accuracy of 0.6618 and $F1$ score of 0.7536. The random forest model had better values of accuracy and $F1$ score than the SVM and the logistic model. The SVM had better values of accuracy and $F1$ score than the logistic model. The non-parametric models, that is the SVM and the Random Forest, performed better than the parametric model, logistic regression model. Arjun *et al.* (2013) fitted a linear kernel SVM model and gausian kernel models to predict flight delays. The linear kernel showed an accuracy of 0.8892 and $F1$ score of 0.576 while the gausian kernel model had an accuracy of 0.8392 and $F1$ score of 0.557 (Arjun *et al.*, 2013). These models had better performance than the SVM model fitted for this study. However, the sample size used in fitting the models was smaller than the sample size that fitted the SVM model for this study. Eisinga (2016) predicted runway allocation using SVM and logistic regression. The two models yielded similar performance with logistic regression performing slightly better. Chow (2018) predicted financial credit risk using logistic regression model, support vector machine and decision trees and compared the models using accuracy and $F1$ score. The logistic regression model showed better performance than the support vector machine and decision tree. This implies that in different prediction problems the fitted models will show varied results.

Table 1: Models Comparison using Train Data

Model	Accuracy	Precision (positive predicted value)	Recall (sensitivity)	$F1$ Score
Logistic Regression Model	0.6618	0.6695	0.8619	0.7536
Support Vector Machine Model	0.6862	0.6832	0.8990	0.7764
Random Forest Model	0.6899	0.6532	0.9658	0.7793

Table2: Models Comparison using Test Data

Model	Accuracy	Precision (positive predicted value)	Recall (sensitivity)	$F1$ Score
Logistic Regression Model	0.6608	0.6721	0.8540	07522
Support Vector Machine Model	0.6828	0.6814	0.9031	0.7767
Random Forest Model	0.6677	0.6538	0.9654	0.7796

Conclusion

This study involved comparison of prediction powers of the logistic regression model, the support vector machine model and the random forest model as used in prediction of flight delays at Jomo Kenyatta International Airport. This was motivated by interest to find out whether non-parametric models perform better than the parametric models in predicion of flight delays. The models were fitted using R software. A good random forest model was achieved a smaller sample size (3000 flights) as compared to the logistic model and the support vector machine (5000 flights). This implied better flexibility of the random forest model in fitting the data. The random forest had better values accuracy and $F1$ score when compared to the SVM and the logistic model. The SVM had better accuracy and $F1$ score than the logistic model. The two non-parametric models (SVM and Random Forest) performed better than the parametric model (logistic model). This was attributed to their flexibility in fitting the data since the do not assume an initial form of distribution of the response.

References

Arjoun, M., Aaron, N. & Kenny, N. (2013). *Predicting Flight on-Time Performance.* Retrieved from https://pdfs.semanticscholar.org/ad5f/f1f4170218ac8b816e 940e75a5e5f941fd42.pdf.

Burgauer, D. & Peters, P. (2000). Airline Flight Delays and Flight Schedule Padding. *Science Investigative Report*, University of Pennsylvania, Philadephia.

Chow, J. C. (2018). Analysis of Financial Credit Risk Using Machine Learning. *arXiv preprint arXiv*:1802.05326.

Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and other Kernel Based Learning Methods*. Cambridge University Press.

David, M.W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies. 2 (1): 37–63.*

Eisinga, K. (2016*). Predicting Runway Allocation with Support Vector Machine and Logistic Regression*. Doctoral Dissertation in Science, Tilburg University, Netherlands.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters. 27 (8): 861–874.*

Finlay, S. (2014). Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods (1st edition.). Basingstoke: Palgrave Macmillan. p. 237. ISBN 1137379278.

Geisser, S. (2016). *Predictive Inference*. Retrieved from *https://www. routledge.com /Predictive-Inference/Geisser/p/book/9780203742310.*

Hao L., Hansen L., Zhang Y. & Post J. (2014): Two Ways of Estimating the Delay Impact of New York Airports. *Transportation Research Part E: Logistics and Transportation Review*

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112). New York: Springer.

Kalliguddi, A.M. & Leboulluec AK (2017). Predictive Modeling of Aircraft Flight Delay. *Universal Journal of Management* 5(10): 485-491.

Lunt, M. (2013). *Introduction to Statistical Modelling*: Linear Regression. Rheumatology, 54(7), 1137-1140.

Meek, C., Thiesson, B.; Heckerman, D. (2002). "The Learning-Curve Sampling Method Applied to Model-Based Clustering". *Journal of Machine Learning Research. 2 (3): 397*

Mistry, P., Neagu, D., Trundle, P. R. & Vessey, J. D. (2016). Using Random Forest and Decision Tree Models for a New Vehicle Prediction Approach in Computational Toxicology. *Soft Computing*, 20(8), 2967-2979.

Mueller, E.R. & Chatterji. G.B. (2002). Analysis of Aircraft Arrival and Departure Delay Equilibrium. *University of California, Berkeley*.

Olson, M., David, L. & Delen, D. (2008). *Advanced Data Mining Techniques*. Springer, 1st edition. ISBN 3-540-76916-1

Pan, Y., Jiang, J., Wang, R. & Cao, H. (2008). Advantages of Support Vector Machine in QSPR Studies for Predicting Auto-ignition Temperatures of Organic Compounds. *Chemometrics and Intelligent Laboratory Systems,* 92(2), 169-178.

Qin Q. & Yu H. (2014). A Statistical Analysis on the Periodicity of Flight Delay Rate of the Airports in the USA. *Advances in Transportation Studies*.

Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and Prediction of Air Traffic Delays. *Transportation Research Part C: Emerging Technologies*, 44, 231-241.

Sheskin, D. J. (2011). Parametric Versus Nonparametric Tests. *In International Encyclopedia of Statistical Science* (pp. 1051-1052). Springer, Berlin, Heidelberg.

Sternberg, A., Soares, J., Carvalho, D. & Ogasawara, E. (2017). A Review on Flight Delay Prediction. *arXiv preprint arXiv*:1703.06118.

Sui, H., Khoo, C. & Chan, S. (2003). Sentiment Classification of Product Reviews Using SVM and Decision Tree Induction. *Advances in Classification Research Online*, *14*(1), 42-52

Tripathi, G. & Naganna, S. (2015). Feature Selection and Classification Approach for Sentiment Analysis. Machine Learning and Applications: *An International Journal*, 2 (2), 1-16.

Waljee, A.K., Higgins, P.D. & Singal, A.G. (2014). A Primer on Predictive Models. *Clinical and Translational Gastroenterology,* 5(1), e44