LITERATURE SURVEY

The main concern of the researchers and analysts is to predict the reasons for flight delays and for that they have put in their efforts on collecting data about flight and the weather. Mohamed et al. [2] have studied the pattern of arrival delay for non-stop domestic flights at the Orlando International Airport. They focused primarily on the cyclic variations that happen in the air travel demand and the weather at that particular airport.

In Shervin et al.'s work [3], their motive of research is to propose an approach that improves the operational performance without hampering or effecting the planned cost.

Adrian et al. [4] have created a data mining model which enables the flight delays by observing the weather conditions. They have used WEKA and R to build their models by selecting different classifiers and choosing the one with the best results. They have used different machine learning techniques like Naïve Bayes and Linear Discriminant Analysis classifier.

Choi et al. [5] have focused on overcoming the effects of the data imbalancing caused during data training. They have used techniques like Decision Trees, AdaBoost, and K-Nearest Neighbors for predicting individual flight delays.A binary classification was performed by the model to predict the scheduled flight delay.

Schaefer et al. [6] have made Detailed Policy Assessment Tool (DPAT) that is used to stimulate the minor changes in the flight delay caused by the weather changes.

Bing Liu [7] has done a sentiment analysis and opinion mining that analyzes people's opinions, sentiments, and studies their behavior. The output of

the research is a feature-based opinion summary which is also known as sentiment classification.

Using techniques such as Natural Language Processing, Naïve Bayes, and Support Vector Machine, researchers built algorithms for analysis that helped them in extracting features in the model. Most of them focused on predicting overall flight delays. Our research concentrated mainly on predicting flight delays for a particular airport over a specific period of time. First, we used a regression model to examine the significance of each feature and then, a feature selection approach to examine the impact of feature combination. These two techniques determined the features to retain in the model. Instead of using the whole set, we sampled 5,000 records at a time to run through different machine learning models. The machine learning models implemented here were Random Forest classifier and Support Vector Machine (SVM) classifier. Further, we applied an approach 5 called One-Hot-Encoder to create a variant of the model for evaluating potential prediction performance.

REFRENCE :

[1] A. B. Guy, "Flight delays cost $32.9 billion, passengers foot half the bill". [Online] Available : https://news.berkeley.edu/2010/10/18/flight_delays/3/. [Accessed on June 2017].

[2] M. Abdel-Aty, C. Lee, Y. Bai, X. Li and M. Michalak, "Detecting periodic patterns of arrival delay", Journal of Air Transport Management,, Volume 13(6), pp. 355– 361, November, 2007.

[3] S. AhmadBeygi, A. Cohn and M. Lapp, "Decreasing Airline Delay Propagation By Re-Allocating Scheduled Slack", Annual Conference, Boston, 2008.

[4] A. A. Simmons, "Flight Delay Forecast due to Weather Using Data Mining", M.S. Disseration, University of the Basque Country, Department of Computer Science, 2015.

[5] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms", Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th, Sacramento, CA, USA, 2016.

[6] L. Schaefer and D. Millner, "Flight Delay Propagation Analysis With The Detailed Policy Assessment Tool", Man and Cybernetics Conference, Tucson, AZ, 2001.

[7] B. Liu "Sentiment Analysis and Opinion Mining Synthesis", Morgan & Claypool Publishers, p. 167, 2012. 55

[8] Statistical Computing Statistical Graphics. [Online]. Available: http://statcomputing.org/dataexpo/2009/the-data.html. [Accessed on April 2017].

[9] FAA Operations & Performance Data. [Online]. Available: https://aspm.faa.gov/.[Accessed on April 2017].

[10] B. Bailey, "Data Cleaning 101". [Online]. Available: https://towardsdatascience.com/data-cleaning-101-948d22a92e4. [Accessed on March 2018].

[11] P. Panov, L. Soldatova and S. Džeroski, " OntoDM-KDD: Ontology for Representing the Knowledge Discovery Process", Discovery Science 2013, Volume 8140, pp. 126-140, 2013.

[12] Bureau of Transportation Statistics. [Online]. Available: https://www.transtats.bts.gov/carriers.asp. [Accessed on 2 April 2017].

[13] How to Predict Yes/No Outcomes Using Logistic Regression. [Online]. Available: https://blog.cleaarbrain.com/posts/how-to-predict-yesno-outcomes-using-logisticregression [Accessed on 3 Feubrary 2018].

[14] S. Polamuri, "How The Random Forest Algorithm Works In Machine Learning". [Online]. Available: https://medium.com/@Synced/how-random-forest-algorithmworks-in-machine-learning-3c0fe15b6674. [Accessed January 2018].

[15] S. Ray, "Understanding Support Vector Machine algorithm". [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vectormachine-example-code/.[Accessed November 2017].

[16] OneHotEncoder. [Online]. Available: http://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. [Accessed on March 2018].

[17] R. Vasudev, "Why and When do you have to use OneHotEncoder?".[Online]. Available: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-doyou-have-to-use-it-e3c6186d008f. [Accessed on March 2018].

[18] Twitter API Twitter. [Online]. Available: https://developer.twitter.com/en/docs.

[19] S. Loria , "TextBlob: Simplified Text Processing", 2016. [Online]. Available: http://textblob.readthedocs.io/en/dev/ [Accessed on December 12, 2017].

[20] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data," Columbia University, New York, December, 2011.

[21] V. A. Kharde and S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (0975 – 8887), Volume 139, no.11, p.11, April 2016.