IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Cancellation Prediction for Flight Bookings using Machine Learning

Ahlam Ansari[1], Salim Mapkar [2], Ashad Shaikh[3*], Maaz Khan[4]

[1,2,3,4] Dept. of Computer Engineering, M.H. Saboo Siddik College of Engineering, Mumbai, India

[*]*Corresponding Author: shaikhashad35@gmail.com Tel.: +91-97687-10696*

DOI: https://doi.org/10.26438/ijcse/v7i3.319321 | Available online at: www.ijcseonline.org

Accepted: 23/Mar/2019, Published: 31/Mar/2019

***Abstract*—**To generate revenue for any service-based industry, selling the right product to the right customer at a right time is the key aspect. Airline industry is an example of such an industry which could get benefit from knowing the right type of customers. This type of customers can be found out by analyzing behavioral patterns over a brief period of time. . Cancellation of flight ticket bookings is an interesting aspect from the perspective of Airline industries. If there is a system available which can predict about customer's cancellation of booking then it can be exploited for huge profits and identifying customers which might possibly cancel their bookings is one of the many tasks that can be achieved by leveraging Data Analytics and Machine Learning techniques. Our goal is to design and implement a Classification model which will predict cancellation of ticket booked. We intend to achieve this goal by analyzing ticket booking data of a domestic Indian airline with the help of data analysis techniques to find some interesting patterns in the data. The predicted output will help to scale down the loss of Airline industries.

***Keywords*—**Cancellation prediction, Flight data analytics, Machine learning.

I. INTRODUCTION

Introduction Data Analytics and Machine Learning is an emerging technology in computer science. Data Analytics enables us to use innovative tools and techniques to analyse and study data over a brief amount of time and Machine Learning algorithms allows us to train Classification models on this data. The data that we will handle in this project is of domestic Indian Airline and we focus on one key aspect of our data: Cancellation of bookings.

This paper proposes a system based on Machine Learning and Data Analysis techniques to extract important features from the dataset using preprocessing tools available in Pandas[1] library and to train and hyper tune a machine learning model using the library Scikit-learn[2] in python which can then be implemented in the airline industry for predicting whether a booking will be cancelled with the goal of generating more revenue for the Airlines.

II. RELATED WORK

Booking cancellations is an important aspect for generating revenue in airline industry, but determining cancellation is a complex task due to the various features and other factors. The reasons for cancellation include many factors such as emergency medical situation, bad weather or some last moment changes in the schedule. Sometimes, airline industries also allow over bookings even though there is a risk of the number of passengers allotted to a certain flight being greater than the capacity of the airlines. But however, if a booking is cancelled at the last moment then the airlines might also incur a loss due to vacant seat, hence there must be balance between both the factors. In current scenario, when a cancellation is done, it is completely sudden without any prior knowledge. Hence, it makes managing such situations a cumbersome task. The process of existing system is shown in Figure 1.

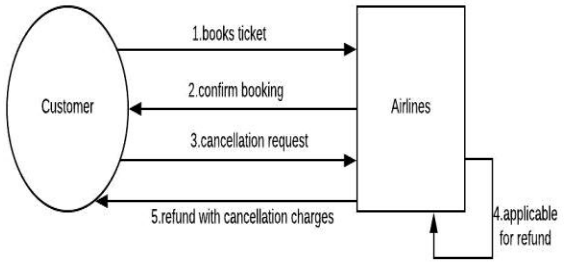

Figure 1. Block Diagram of existing system

Petraru[3] infers that cancellation rate forecasting in combination with a overbooking policy can increase the revenue of airlines, Cancellation prediction can result the gain revenue by 0.12% even without overbooking which

marks the cancellation prediction a very important factor in generating revenue for airlines system.

There are roughly two different approaches for cancellation models: first, forecasting of cancellation rates and second, classify each reservation individually, so-called Passenger Name Record (PNR) approach. Classifying cancellations based on PNR approach is quite popular in the airline industry according to Petraru [3]. PNR approach tries to learn patterns from record-based data that contains information about the bookings and passengers. Antnio, Almeida and Nunes [4] used similar PNR based approach in hospitality industry, to train a decision tree with a result of 98.6% accuracy on cancellation prediction model.

III. METHODOLOGY

Cancellation of bookings can either be a boon or bane for the airlines. On every cancellation of booking, Airline faces the risk of incurring losses due to flight taking off with a vacant seat. However, this can even be exploited for generating more revenue if the possibility of a booking being cancelled is known beforehand. We propose a solution of predicting cancellation of bookings through Machine Learning and Data Analysis techniques. Customer's behavioral patterns of over a year will be analyzed using the flight booking dataset and then a Prediction Classifier will be trained using different Machine Learning algorithms for finding a optimal classifier which can predict the cancellations with desired accuracy. Figure 2. shows the proposed system.

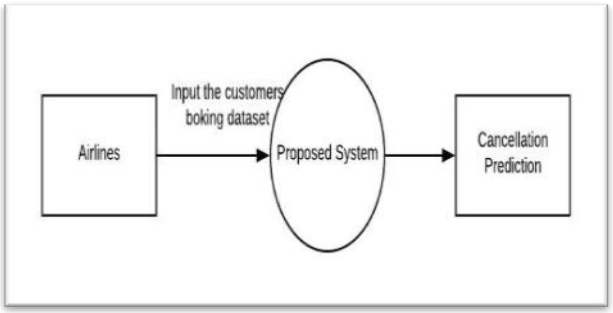

Figure 2. Block Diagram of proposed system

The dataset that will be used is of domestic Indian Airlines. It consists of features related to both passengers and booking data. Examples of features that are booking related are booking date, boarding date, price and via which channel the reservation has been made. Features related to passengers are number of passengers, nationality, type of passenger (adult or child) and more.

Even though quality of the data is high, but for good accuracy, it requires some feature selection. According to Howbert [5], a selection can be beneficial if redundant, irrelevant or noisy features are removed in order to speed up learning process of the model, enhance generalization and to alleviate the curse of dimensionality. New features can also be engineered for boosting the performance of chosen model. Howbert [5] states that well-conceived new features can sometimes capture the important information in a dataset much more effective than the original features. The available data set is given as in Table 1.

Table 1. Columns in dataset

PNR	Book date	Charge date	From	To	.....	Charge code	Status

A. Proposed Algorithm

Following are the steps of the proposed algorithm:

Step 1: Start
Step 2: Input Data of Customer bookings
Step 3: Preprocessing and cleaning the data
Step 4: Feature extraction from the data
Step 5: Train different model using Training data
Step 6: Evaluate the model using testing data
Step 7: Compare accuracy of the model.
Step 8: If results are Satisfactory then go to Step 9 else go to Step 5
Step 9: Feed the new test data to the model.
Step 10: Predict the outcome
Step 11: End

Figure 3. shows the flowchart of the proposed system in which we will train different models using the available dataset. The model with the highest accuracy will be endure. Before training the model, the dataset need to be preprocessed with feature extraction.

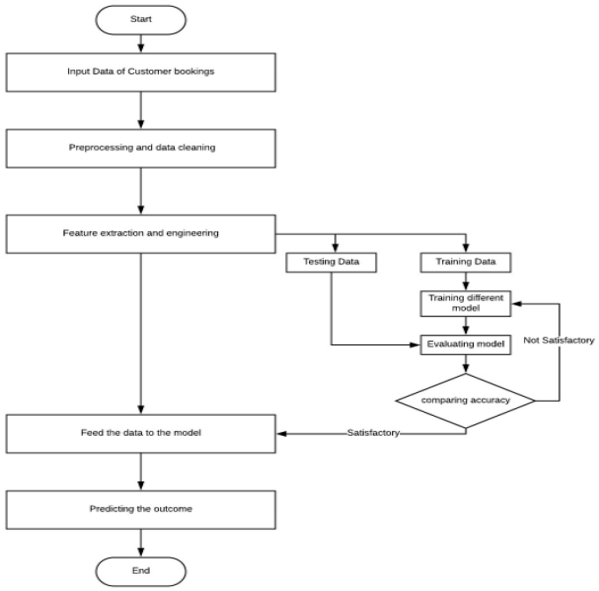

Figure 3. Flowchart of proposed system

IV. RESULTS AND DISCUSSION

We used the entries of single day file as our dataset and performed preprocessing and feature engineering on the dataset. After extracting the required columns in a suitable format, we trained different models on the dataset whose results are specified in Table 1.

Table 2. Comparison of different models.

Classifier(s)	Precision	Recall	F1 score	ROC
Logistic Regression	0.9766	0.7746	0.8640	0.8867
Decision Trees	0.9716	0.9506	0.9609	0.9743
Random Forest	1.0000	0.9043	0.9497	0.9521
Gradient Boosting	0.9743	0.9382	0.9559	0.9683

V. CONCLUSION AND FUTURE SCOPE

We tried to propose a solution to the problem of Cancellation of bookings faced by Airlines Industry by implementing a Cancellation Prediction Classifier by training different Machine Learning models over flight data by analysing them through Machine Learning. Currently, we trained different models for data of a single day as a dry run and compared the results based on different metrics. We aim to scale the project by training different models over 1 year of data and the models will be then compared based on a chosen metrics. The available dataset contains more than 5 million of rows for a single month. Hence to process the data of whole 1 year, it will be efficient to use big data tools and framework such as Apache spark which is an implementation of Resilient Distributed Datasets (RDD) introduced by Zaharia et al. [6] in 2012. Spark engine supports multiple libraries for training machine learning models and the most notable ones are: Spark MLLib and XGBoost. MLLib is Spark's open-source distributed machine learning library which according to Meng et al[7], provides efficient functionality for a wide range of learning settings whereas XGBoost is a scalable end-to-end tree boosting system as described by Tianqi Chen and Carlos Guestrin[8], who concluded that XGBoost is able to solve real-world scale problems using a minimal amount of resources and it is also supported in multiple languages and framework including Python and Spark. More meaningful insights can be obtained by using these tools over a dataset of long time period.

REFERENCES

[1] W. McKinney, pandas: a python data analysis library, http://pandas.sourceforge.net [scipy2010].

[2] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Edouard Duchesna: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011).

[3] O. Petraru: Airline passenger cancellations: modeling, forecasting and impacts on revenue management, Massachusetts Institute of Technology, 2016.

[4] 2. N. Antnio, A. Almeida, L. Nunes: Predicting hotel booking cancellations to decrease uncertainty and increase revenue, Tourism & Management Studies, 2017.

[5] 3. J. Howbert: Introduction to Machine Learning, University of Washington Bothell,2012.

[6] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. In USENIX Symposium on Networked Systems Design and Implementation, 2012.

[7] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. MLlib: Machine learning in apache spark. Journal of Machine Learning Research, 17(34):1–7, 2016.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016

Authors Profile

Prof. Ahlam Ansari Works as an Assistant Professor at M.H. Saboo Siddik College of Engineering in Computer Engineering Department. Qualifications: -. Also, having a teaching experience of more than 10 years in this institute.

Mr. Salim Mapkar currently pursuing Computer Engineering from M.H. Saboo Siddik College of Engineering and belongs to Department of Computer Engineering.

Mr. Ashad Shaikh. currently pursuing Computer Engineering from M.H. Saboo Siddik College of Engineering and belongs to Department of Computer Engineering..

Mr. Maaz Khan currently pursuing Computer Engineering from M.H. Saboo Siddik College of Engineering and belongs to Department of Computer Engineering.

