Airfare Prediction using XGBoost and Deep Learning

Ashkan Sharabiani
Adjunct Professor - UIC
ashara2@uic.edu

Akash Tayal
661488456
atayal3@uic.edu

Amit Yadav
670304917
ayadav8@uic.edu

Arnab Mukherjee
675789086
amukhe27@uic.edu

Darshan MA
664760377
dmysor2@uic.edu

Mayur Mishra
658522696
mmishr5@uic.edu

***Abstract*— The advantages of travelling by airplanes have increased overtime. The industry tries to make the ticket fare reasonable as well as to make profit out of it. Airline industry has lots of dynamic factors affecting them in a day today operation. It's one of the highly sophisticated industry which aims at making revenue. The purpose of this study is to better analyze the features that affect airfare and develop and tune models to predict the airfare well in advance. We used XGBoost and Keras with a Tensorflow backend Neural Network, two state of the art prediction models for our study and used various machine learning tasks to achieve the best performance for our task.**

Keywords— XGBoost, Neural Network, Prediction model, airfare,

I. INTRODUCTION

Since the deregulation of the airline industry in the 1907's, researchers have investigated the mystery of airfares and how airlines determine them. Many factors play a role in how airlines go about determining airfares. These factors include various elements found to be significant over the years such as number of competitors, number of passengers, distance between origin and destination, etc. While many studies look at the role of hub in airline pricing, few studies concentrate on trying to determine some of the factors playing a role in airfare pricing in hub-to-hub markets, as this paper attempts. Hub and Spoke operation frameworks, which permit bearers to serve more urban communities in a seemingly more proficient way, are the favored system structure of the aircraft industry. Past research by Borenstein (1989) demonstrates that while hubs are effective operating devices for aircrafts in terms of the quantity of various markets the aircraft can serve, they are detrimental for consumers because the airlines become isolated from competition when they have a monopoly at their hubs His research concludes that when a carrier has an overwhelming position in an airplane terminal (for the most part at a hub), that carrier will charge higher fares than in whatever remains of its framework. [2]

This project aimed at developing a model using two different approaches for Airline Price prediction. Model was developed in python programming language. Our work benefits the airline industry to increase their revenue. We aimed to predict the price to find the business advantages for the airlines as well. In this project we studied the ticket price varying over time and many additional factors affecting them. This model would help airlines to prioritize their actions towards the coach class-fares.

The remainder of this paper is organized as follows: Section 2 speaks about the Literature Review; Section 3 shows the Data Preprocessing followed by Data Visualization in Section 4. Section 5 shows how the Modelling was done; Section 6 presents a detailed reporting of the Results obtained; Section 7 draws the Conclusions; followed by References in the last section.

II. LITERATURE REVIEW

Tianqi Chen (2016) speaks about how Tree boosting is one of the most important and widely used machine learning models. XGBoost takes a top down approach, by building a scalable tree boosting system on top of a few primitives for which the implementation can be easily replaced. He also shows that distributed XGBoost can easily handle billion scale dataset, and gives near linear speedup with more machines. Alekseev and Seixas (2009) show that in time series analysis it is assumed that the data have a systematic pattern – usually a set of identifiable deterministic components (trend and seasonality) and random noise.

Nelson et al. (1997) investigated the ability of the neural network in 'learning' the seasonally of the series, using the time series of competition M and found that the neural networks were unable 'to learn' the seasonality and their forecasting ability was harmed, although they still concluded that the networks can obtain benefits of the non-seasonal data because best linear unbiased estimators are robust for such cases.

The welfare effects of fare discrimination across periods in time were measured in the paper (Lazarev, 2013). Lazarev developed an empiricalmodel of optimal prices and found out that due to intertemporal price discrimination airlines can earn approximately 90% of the profit.

In the paper (Etzioni, 2003) the multistrategy data mining algorithm was presented to find the optimal time for purchasing (for the last 21 days prior to departure).

III. DATA PRE-PROCESSING

According to Techopedia, Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

The dataset was first checked for missing values and outliers to better analyze the data. The dataset after preliminary analysis was found to have no missing values or outliers.

In many practical Data Science activities, the data set will contain categorical variables. These variables are typically stored as text values which represent various traits. Regardless of what the value is used for, the challenge is determining how to use this data in the analysis. Many machine learning algorithms can support categorical values without further manipulation but there are many more algorithms that do not. XGBoost is one example of such model. Therefore, we were faced with the challenge of figuring out how to turn these text attributes into numerical values for further processing. Since we are dealing with XGBoost as our prediction model, we could not use the categorical variables like Origin, Destination etc.

Encoding categorical variables is an important step in the data science process. Sklearn's LabelEncoder module finds all classes and assigns each a numeric id starting from 0. This means that whatever your class representations are in the original data set, you now have a simple consistent way to represent each. We thus converted the categorial features in our data set to numerical values using Label Encoding.

We also used Feature Scaling for our Neural Network model by using MinMaxScalar. Min-max normalization is often known as feature scaling where the values of a numeric range of a feature of data, i.e. a property, are reduced to a scale between 0 and 1. It essentially shrinks the range such that the range is now between 0 and 1.

IV. DATA VISUALIZATION

Data visualization is the presentation of data in a pictorial or graphical format and it helps decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments. Its one of the most important Data Science Technique which helps us identify areas that need improvement, predict sales, clarify what factors influence behavior etc.

For our dataset, we initially tried to understand the relationship between the time when the fares were filed with IATA and the months along with the most popular airports to better understand how different factors played a role in increasing the fare. We tried to the explore the connection between different features from the correlation plots also.

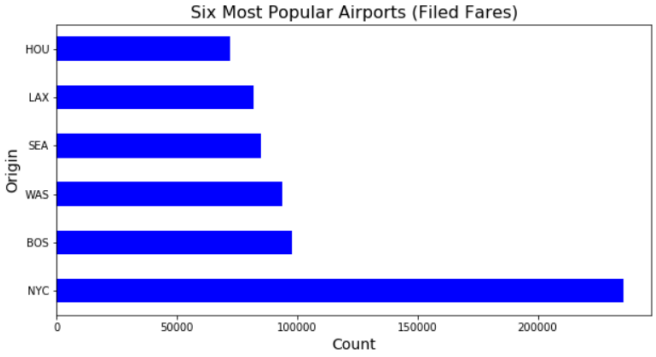

Fig 1. Most popular airports in terms of filed fares

The above graph shows that New York City faced the most price fluctuations owing to its popularity as compared to all the other destinations in US.

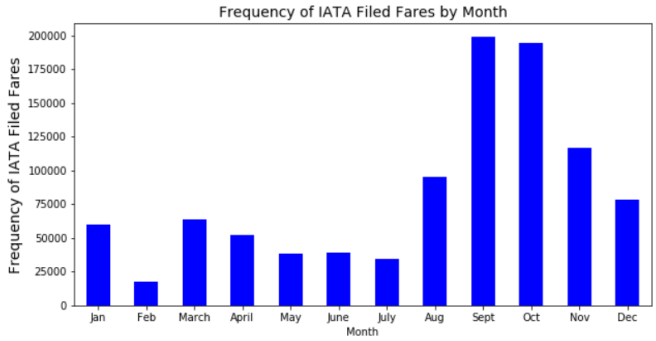

Fig 2. Frequency of IATA filed fares v/s Month

The above graph shows that as we move towards the holiday, price fluctuations keep increasing as airlines look to increase their profits because of the high demand in the winter months. Thus, from August to October, we can see that the airline company keep changing the price for the holiday season according to different trends.

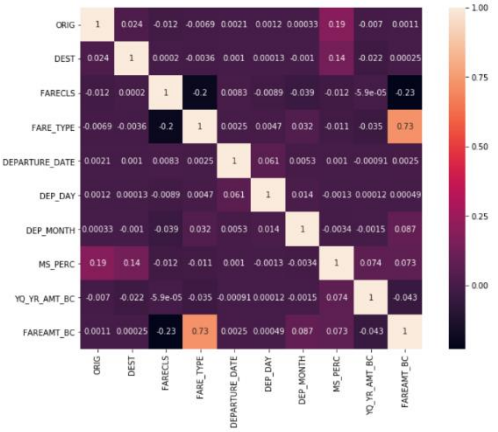

Fig3. Correlation Plot

The above correlation plot shows the positive and negative relationship between different features. We can see that fare type with fare amount and fare type with market share have a high correlation.

V. MODELLING

After our literature review, we zeroed down two models for our comparison, XGBoost and Neural Network. We used the Keras Library (Tensorflow backend) for modelling purposes in the form of a Sequential Model.

A. Model

Among the various machine learning methods used in practice, gradient tree boosting is one technique that shines in many applications. Tree boosting has been shown to give state-of-the-art results on many standard classification benchmarks. Variants of the model have been applied to problems such as classification and ranking.

In this paper, we use XGBoost as our prediction model, a scalable machine learning system for tree boosting. The system is available as an open source package. The impact of this boosting method has been widely recognized in many machine learning and data mining challenges. The most important factor behind the success of XGBoost is its scalability in all scenarios. Tianqi Chen (2016) spoke about how the system runs more than ten times faster than existing popular solutions on a single machine and scales to billons of examples in distributed or memory-limited settings. Parallel and distributed computing makes learning faster which enables quicker model exploration. More importantly, XGBoost exploits out-of-core computation and enables data scientists to process millions of examples on a desktop. [6]

Deep learning is basically an advanced type of machine learning that imitates the workings of the human brain in processing data and creating patterns for use in decision making. It is good at combining contrasting type of data which are distributed over time and space. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. [5]

We used the feed forward ANN formed from thousands of simple processing units, connected in parallel in this paper to compare its performance with a boosting technique. With a differentiable squashing function usually the sigmoidal function, the feed-forward neural network is a network of perceptron's.

B. Hyperparameter Tuning

Hyperparameter tuning works by running multiple *trials* in a single training job. Each trial is a complete execution of your training application with values for your chosen hyperparameters set within limits you specify. The accuracy of the model, as calculated from an evaluation pass, is a common metric. The metric must be a numeric value, and you can specify whether you want to tune your model to maximize or minimize your metric. When you start a job with hyperparameter tuning, you establish the name of your hyperparameter metric.

We used the most popular approach, Grid Search for tuning of our XGBoost hyperparameters. In Grid Search, we try of different boosting hyperparameters, so that it forms a grid of configurations and train the algorithm accordingly choosing the configuration that gives the best performance. We specified the bounds and steps between values of the hyperparameters so that it forms a grid of configurations. We started with a limited grid with relatively large steps between parameters values, then extended to make the grid finer at the best configuration. The evaluation metric for the grid search was set to Root Mean Square Loss.

After running the grid search for various combinations, we found the optimal number of estimators to be 500, colsample and subsample ratio to be 0.8, maximum depth of the tree as 500, minimum child weight to be 5 and learning as 0.3.

C. Layer Definition

The architecture of our neural network was not very complex and used only one hidden layer followed by an output layer to predict the price. A Rectified Linear Unit (ReLu) activation function was used at each of the first two layers. The output space, number of neurons, of every layer was 600, 400, and 1 for layers 1, 2, and 3 respectively. The layer types, in Keras terms, is Dense. A Dense layer means a fully connected layer was used with no connections present between neurons within a layer. In other words, very neuron in each layer is fully connected to all neurons in the successive layer. We used Adam as our optimization algorithm for the output layer to update network weights. The name Adam is derived from the adaptive moment estimation. Specifically, the algorithm calculates an exponential moving average of the gradient and the squared gradient and the beta parameters control the decay rates of these moving averages. [7]

D. Parameters

For our deep learning model, we also included epochs and batch size. An epoch is a situation where the model sees, in a sense, all the instances of the dataset under processing. The number of epochs was set to 500 to increase the model's familiarity with the data without creating an overfitting scenario. For the purposes of increasing the computational efficiency of the algorithm, the dataset was batched before being passed into the model. A batch size of 50 was used to prevent overfitting since the model sees every data point as its own rule for prediction.

E. NVIDIA GPU

NVIDIA GPUs power millions of desktops, notebooks, and workstations around the world, accelerating computationally-intensive tasks for scientists, and researcher. We used the CUDA platform as it included parallel computing extensions, powerful drop-in accelerated libraries and cloud based computing. We decided to use GPU support to reduce the computational time as our dataset was quite large. Neural Network with an epoch size of 500 took about 12 hours for just one simulation run whereas XGBoost with Hyperparameters tuning for various combinations took about 6 hours. GPU support was a big relief for our case as it reduced computation time on a large scale and helped us run many simulations to further improve the MAE and MSE

VI. Results

This section compares both the models perform and then selects the best one for our case. We compared the MSE and MAE performance of both the models and considered their individual computation time to select the best fit.

	XGBoost	Neural Network
MSE	2100	3000
MAE	32.8	40

Fig 4. Comparison of performance between models

As we can see from the above figure, XGBoost out performs the Neural network in predicting airfare. MAE shows that the prediction is just short of $32.8 for the XGBoost model and $40 for the Neural Network model. This is great from airlines point of view as they don't have to manually change the airfares. These models take into consideration the monthly trends and the most popular airport in terms of price fluctuations.

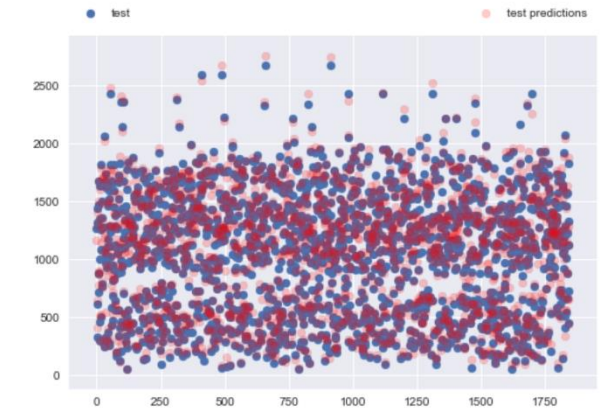

Fig 5. Scatter Plot between Test set and Predictions (Neural Network)

The above figure shows the scatter plot between our test dataset and predictions. We can observe that the model is quite efficient in terms of predicting the airfare. The model is a great fit and both the test and predictions coincide on the scatter plot.

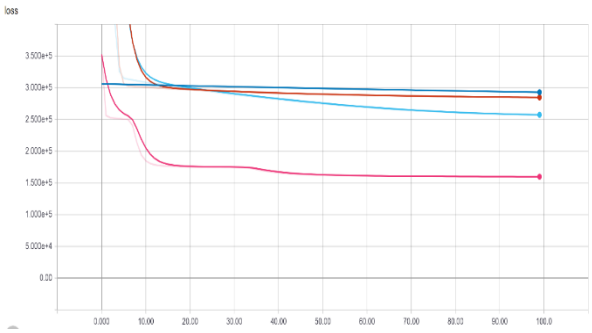

Fig 6. Loss Function plot - Tensorflow

The above figure shows how the loss function reduced as we kept varying parameters like the number of epochs, batch size, learning rate etc. The final red line gave us the best performance when we used 500 epochs with a batch size of 50.

XGBoost has an inbuilt function called feature importance which tells us how each feature contributed towards predicting the target variable. This is very important as it allows the industry to control certain features to impact the final price.

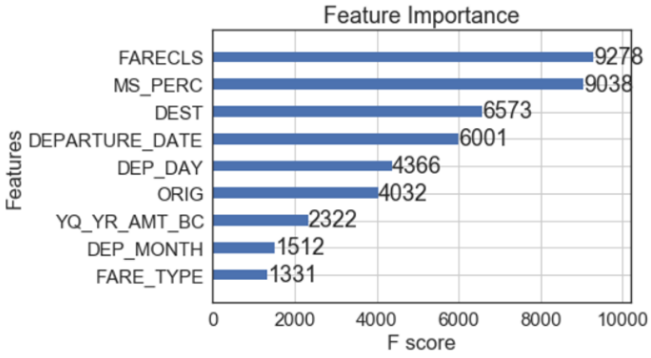

Fig 6. XGBoost Feature Importance

Thus, we can see that Fare Class, Market share of the airline company in that country, Destination and Departure date as the most driving factors in the prediction of airfare. This is in accordance with our basic understanding of the how the airlines industry price the tickets. Thus, the airline company can use these features to better predict the price and leverage these features to increase their profits.

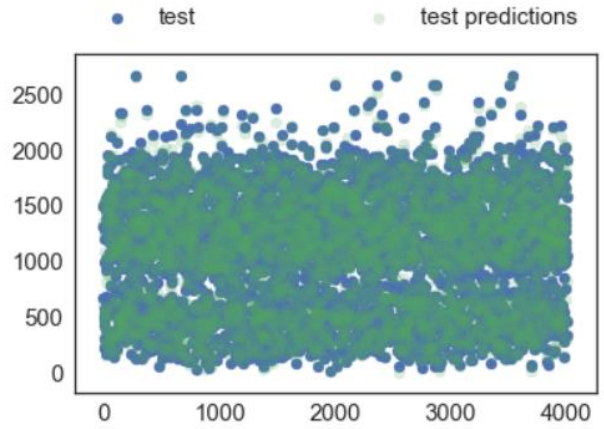

Fig 8. Scatter plot between Test set and Predictions (XGBoost)

Thus, we can conclude from the scatter plot above that XGBoost is a better model as all the points fit better in this system as compared to Neural Network. The 'green' predictors coincide well with the test data set with no overfitting issues.

Computationally also, XGBoost took 6 hours lesser to compile and run thus proving to be very efficient for our case.

VII. Conclusion

In this paper we used XGBoost, a scalable tree boosting system and a feed forward neural network for airfare price analysis. Our analysis & results show XGBoost as the best prediction model with much lower errors and computational time. Moreover, it indicates Fare class & Market share of the

airline company being the most important features affecting the prices. By using these insights, the airline company can better control their prices by focusing on business strategies to increase market share by adjusting business class fares.

VIII. REFERENCES

[1] K.P.G. Alekseev, J.M. Seixas*(2009), A multivariate neural forecasting modeling for air transport –Preprocessed by decomposition, / Journal of Air Transport Management 15 (2009) 212–216.

[2] Timothy M. Vowles*, Airfare pricing determinants in hub-to-hub markets, Journal of Transport Geography 14 (2006) 15–22.

[3] Anastasia Lantseva, KseniaMukhina, Anna Nikishova, Sergey Ivanov and Konstantin Knyazkov, Data-driven Modeling of Airlines Pricing, YSC 2015,Volume 66, 2015, Pages 267–276.

[4] Benny Mantin*, BonwooKoo Weekend effect in airfare pricing, Journal of Air Transport Management 16 (2010) 48–50.

[5] R. D. Hof, "MIT Technology Review," [Online]. Available: https://www.technologyreview.com/s/513696/deep-learning/. [Accessed 20 November 2017]

[6] TianqiChen, Carlos GuestrinXGBoost: A Scalable Tree Boosting System, arXiv:submit/1502704 [cs.LG] 9 Mar 2016

[7] K. Documentation, "Core Layers," October 2017 Available: https://keras.io/layers/core/. [Accessed October 2017].

[8] Nelson, M., Hil, T., Remus, W., O'Connor, M., 1997. Time Series Forecasting usingNeural Networks: Should the Data be Deseasonalized First? Working Paper. University of Hawaii.

[9] Etzioni, O. (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price.Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 119-128.

[10] Lazarev, J. (2013). The welfare effects of intertemporal price discrimination: an empirical analysis of airline pricing in US monopoly markets.