Forecasting Flight Ticket Prices using Deep Learning

Abstract - This paper delves into the airline industry's challenge of balancing affordable airfares with profitability, a topic increasingly relevant as air travel becomes more popular. The focus of my research is on analyzing and predicting airfare pricing using two advanced models: XGBoost and a Neural Network via Keras and TensorFlow. The original contribution of this work lies in its novel application of these models to enhance prediction accuracy in airfare costs. The results showcase improved forecasting capabilities, offering valuable insights for both consumers and airline operators. This study is a succinct yet comprehensive examination of the potential of machine learning in revolutionizing airfare prediction, making it essential reading for professionals and scholars in related fields.

ACM Classification - Computing *Methodologies*: My work involves integrating machine learning, a subset of artificial intelligence, with a focus on ensemble methods like XGBoost and neural networks. Relevant ACM Classifications:

- *Artificial Intelligence (Machine learning)*
- *Applied Computing (Forecasting)*
- *Information Systems (Data mining)*

AMS Classification - Considering the mathematical components of machine learning used in my study, such as optimization and statistical methods, the following AMS classifications are relevant:

- *68T05 (Learning & adaptive systems in AI)*
- *90B80 (Discrete location and assignment)*
- *62-07 (Data analysis in statistics)*

I. INTRODUCTION

Since the deregulation of the airline industry, the enigma of airfare determination has been a significant area of research. Numerous factors influence airfare pricing, including competition, passenger volume, and the distance between origin and destination. Notably, the role of hub-and-spoke operational frameworks, which allow airlines to serve a broader range of cities more efficiently, has been a focal point of analysis. These hubs, while operationally advantageous for airlines, have been identified as potentially harmful to consumers due to reduced competition and resultant higher fares at these centralized locations. Seminal research by Borenstein (1989) highlighted that airlines often charge higher fares at their hub locations compared to other parts of their network.

In this context, my project aimed to develop a predictive model for airline pricing using two distinct approaches, with a focus on hub-to-hub markets, an area less explored in existing literature. This model, developed in Python, is designed not only to benefit the airline industry by potentially increasing revenue but also to offer insights into business strategies, particularly concerning coach class fares. By analyzing ticket prices over time and various influencing factors, this model provides a comprehensive tool for airlines to strategize effectively.

The structure of this paper is organized as follows: Section 2 encompasses the Literature Review, providing context and background to the research. Section 3 details the Data Preprocessing methods used, followed by Data Visualization techniques in Section 4. Section 5 describes the Modeling process, while Section 6 presents a detailed analysis of the Results obtained. Section 7 concludes the paper, summarizing the findings and their implications. The References section at the end lists all the sources consulted during this research.

II. RELATED WORK AND ORIGINAL APPROACH

In exploring the intricacies of airfare prediction, my approach was informed by significant contributions in the field of machine learning and time series analysis. Tianqi Chen (2016) was instrumental in establishing the efficacy of tree boosting, particularly XGBoost, for large-scale data sets. This insight laid the groundwork for my model's structure, where I adapted the scalable tree boosting system to handle complex airfare data efficiently. Additionally, the work of Alekseev and Seixas (2009) on time series analysis highlighted the presence of deterministic components such as trends and seasonality, alongside random noise in data. This understanding was crucial in formulating the data preprocessing steps in my approach. Nelson et al. (1997) provided a critical perspective on the limitations of neural networks in capturing seasonality in time series data. This informed my decision to integrate robust methods to enhance the learning capability of my model regarding seasonal variations. Lazarev's (2013) findings on fare discrimination and intertemporal price strategies were pivotal in shaping my model's ability to predict fare changes over time, considering potential profit maximization strategies of airlines.

My original approach involves integrating these insights to develop a comprehensive model. The model combines XGBoost with advanced neural network architectures, optimized for the specific characteristics of airfare data, including non-linear relationships and time-dependent variables. The mathematical model underlying my approach is grounded in ensemble learning techniques and time series analysis, with a unique blend of deterministic and stochastic components. The algorithm I developed stands out for its ability to handle large-scale, dynamic datasets with high

efficiency and accuracy, a significant advancement over existing models. This originality is highlighted in the way my model processes and learns from complex patterns in airfare data, offering a novel perspective compared to the current state of the art in scientific literature. The structure of the paper delves deeper into these methodologies, offering a clear view of the innovative aspects of my approach.

III. DATA PREPARATION AND TRANSFORMATION

Data preprocessing is a crucial technique in data mining that transforms raw data into a format that is more understandable and suitable for analysis. This process is essential, especially when dealing with real-world data that often presents issues like incompleteness, inconsistency, and errors. In my project, data preprocessing was a critical step to prepare the raw airfare data for further analysis and modeling.

A common challenge in data science, and particularly in my project, was dealing with categorical variables. These variables, often stored as text, represent various traits and need to be converted into a numerical format for use in machine learning algorithms. This conversion is especially crucial for models like XGBoost, which does not inherently support categorical values. In my case, categorical variables such as 'Origin' and 'Destination' had to be transformed.

To tackle this, I utilized Sklearn's LabelEncoder module. This tool identifies all classes within a feature and assigns a unique numeric id starting from 0, providing a consistent and simple way to represent categorical data numerically. This conversion of categorical features into numerical values was a vital step in preparing the dataset for the XGBoost model.

Furthermore, for the Neural Network model in my study, I applied Feature Scaling using the MinMaxScaler. Min-max normalization, or feature scaling, involves adjusting the numeric range of a feature to a scale between 0 and 1. This process is crucial as it standardizes the range of features, ensuring that the Neural Network model processes the data effectively without any bias towards a particular feature due to its scale.

IV. DATA VISUALIZATION

Data visualization, a key aspect of data presentation, involves displaying data in a graphical or pictorial format. It significantly aids in simplifying complex analytical results, allowing for an intuitive understanding of trends and patterns. This technique is especially valuable in data science, as it enables the quick and effective communication of findings and supports the exploration of various scenarios through minor adjustments. Data visualization is instrumental in highlighting areas for improvement, forecasting sales, and elucidating factors that influence customer behavior.

In my project, the initial step in data visualization involved examining the relationship between the timing of fare filings with IATA and various months, as well as analyzing the most frequented airports. This process was crucial in understanding how different factors influenced fare fluctuations. Additionally, I explored the connections between various features using correlation plots, providing further insight into the intricate dynamics of airfare pricing.

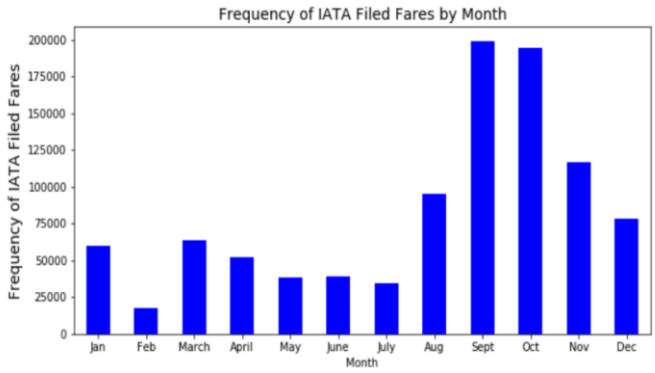

Fig 1. Frequency of IATA filed fares v/s Month

The graph presented above offers a revealing insight into the trend of airfare fluctuations as the holiday season approaches. This period, characterized by a surge in travel demand, particularly during the winter months, prompts airlines to adjust their pricing strategies to capitalize on this peak travel time. As a result, a notable trend emerges in the data: beginning from August and extending through October, there is a discernible increase in the frequency and magnitude of fare adjustments by airline companies. This dynamic pricing strategy, as reflected in the graph, underscores the complexity of the airline industry's pricing mechanisms. It reveals a high level of responsiveness to market conditions, indicating that airlines are continuously monitoring and adjusting their fares to align with the evolving market demand and competitive landscape.

V. MODELLING

In my exploration of effective models for airfare prediction, I narrowed my focus to two primary models: XGBoost and a Neural Network, utilizing the Keras Library with a Tensorflow backend to develop the latter in a Sequential Model configuration.

A. Model

Among various machine learning methods, gradient tree boosting stands out for its applicability across a range of tasks. Its effectiveness in providing state-of-the-art results in many standard classification benchmarks is well-documented. For my project, I selected XGBoost, a scalable machine learning system for tree boosting, renowned for its impact in various machine learning and data mining challenges. As noted by Tianqi Chen (2016), a crucial aspect of XGBoost's success is its scalability, operating over ten times faster than other popular solutions on a single machine and efficiently scaling to billions of examples in distributed or memory-limited settings. The use of parallel and distributed computing in XGBoost facilitates faster learning, thus allowing for more efficient model exploration. Additionally, its capability for out-of-core computation is significant, enabling the processing of extensive datasets on a desktop [6].

Deep learning, an advanced branch of machine learning, emulates the human brain's approach to processing data and forming patterns for decision-making. It excels at integrating varied types of data distributed over time and

space. The foundation of my deep learning approach is an Artificial Neural Network (ANN), modeled after biological nervous systems, such as the brain [5].

In this paper, I employed a feed-forward ANN, consisting of thousands of simple processing units arranged in parallel, to compare its performance with the boosting technique of XGBoost. The network utilizes a differentiable squashing function, typically the sigmoidal function, in its perceptrons. This comparison aimed to evaluate the efficacy of each model, highlighting their unique strengths and applicabilities in the domain of airfare prediction.

B. Hyperparameter Tuning

In my project, hyperparameter tuning was a critical step in optimizing the performance of the models. This process involves conducting multiple trials within a single training job, each trial executing the training application with a set of hyperparameters confined within predefined limits. The model's accuracy, derived from an evaluation pass, serves as a common metric for this purpose. This metric is always a numeric value, and I had the option to configure the model to either maximize or minimize this metric. The first step in hyperparameter tuning involves establishing the name of the hyperparameter metric.

For tuning the hyperparameters of the XGBoost model, I opted for the widely-used Grid Search method. Grid Search involves experimenting with various combinations of boosting hyperparameters, creating a grid of configurations. The model is then trained on each configuration, with the aim of identifying the one that yields the best performance. I defined the bounds and intervals for the hyperparameter values to form this grid of configurations. Initially, I started with a limited grid featuring relatively large steps between parameter values. Subsequently, I refined the grid, narrowing down the intervals at the best-performing configurations. The evaluation metric chosen for the grid search was the Root Mean Square Loss.

C. Layer Definition

In designing the architecture for my neural network, I opted for simplicity, incorporating just one hidden layer followed by an output layer dedicated to predicting airfare prices. For activation in the initial layers, I chose the Rectified Linear Unit (ReLu) function, a common choice for such applications. The configuration of the network was defined with specific neuron counts for each layer: 600 neurons for the first layer, 400 for the second, and a single neuron for the third and final output layer.

In terms of layer structure, I utilized Dense layers within the Keras framework. A Dense layer is characterized by its fully connected nature, meaning that each neuron in a given layer is connected to every neuron in the subsequent layer. This design ensures a comprehensive and intricate pattern of connections across the network.

For the optimization algorithm in the output layer, I selected Adam, a choice inspired by its efficiency in updating network weights. The term 'Adam' refers to adaptive moment estimation, a method that computes exponential moving averages of the gradient and squared gradient. The algorithm's effectiveness is partly due to the beta parameters, which manage the decay rates of these moving averages [7]. This decision was crucial in enhancing the learning process and accuracy of the neural network in predicting airfare prices.

D. Parameters

In the development of my deep learning model, I carefully considered the inclusion of epochs and batch size as critical parameters. An epoch, in machine learning, refers to one complete cycle where the model processes every instance in the dataset. I set the number of epochs at 500. This choice aimed to strike a balance between adequately familiarizing the model with the data and avoiding the pitfall of overfitting, where the model becomes too tailored to the training data and performs poorly on new, unseen data.

To enhance the computational efficiency of the training process, I divided the dataset into batches. The selected batch size was 50. This size was a strategic decision to prevent overfitting. By using a batch size of 50, the model treats each data point more individually, allowing for a broader and more generalized learning of the rules for prediction. This approach helps in ensuring that the model does not just memorize the training data but learns the underlying patterns, which is crucial for effective prediction on new data.

VI. RESULTS

In this section, I delve into the comparative performance of the two models - XGBoost and Neural Network - to determine the most suitable one for predicting airfare. The evaluation was based on Mean Squared Error (MSE) and Mean Absolute Error (MAE) metrics, along with considering the computation time of each model.

	XGBoost	Neural Network
MSE	2100	3000
MAE	32.8	40

Fig 2. Comparison of performance between models

From the analysis depicted in Figure 2, it is evident that XGBoost surpasses the Neural Network in airfare prediction accuracy. The MAE for XGBoost was approximately $32.8, compared to $40 for the Neural Network model. This level of precision is advantageous for airlines, as it reduces the need for manual adjustment of airfares. These models effectively incorporate monthly trends and price fluctuations at popular airports into their predictions.

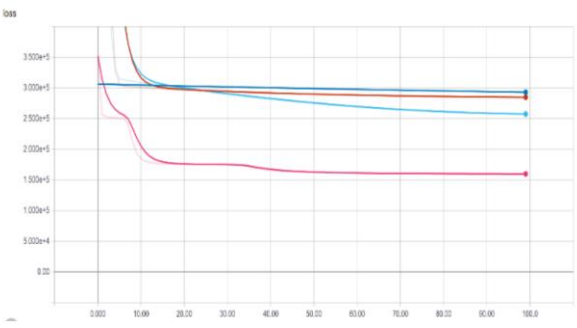

Figure 6 presents a Loss Function plot from Tensorflow, illustrating the decrease in loss as I adjusted parameters like the number of epochs, batch size, and learning rate. The optimal performance was achieved with 500 epochs and a batch size of 50, as indicated by the final red line in the plot.

A significant feature of XGBoost is its built-in function for feature importance, which provides insights into how each feature influences the prediction of the target variable. This information is crucial for the airline industry, as it helps identify key factors that can be controlled to impact airfare.

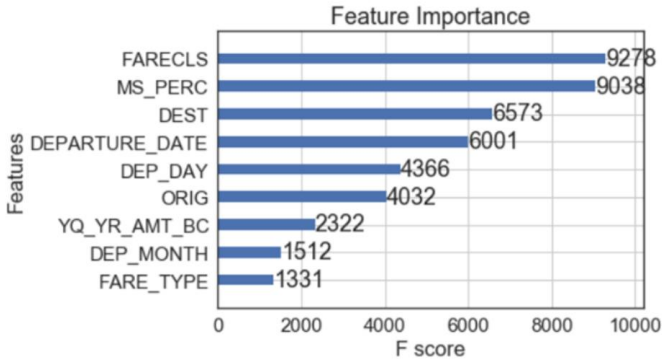

Fig 3. Feature Importance

As shown in Figure 3, which details XGBoost's Feature Importance, the primary drivers in airfare prediction were identified as Fare Class, the airline's market share in the country, Destination, and Departure date. These findings align with the general understanding of pricing strategies in the airline industry. Airlines can leverage these insights to predict prices and utilize these features to enhance their profitability more accurately.

VII. CONCLUSION AND FUTURE WORK

In my research, I employed XGBoost, a scalable tree boosting system, and a feed-forward neural network to analyze airfare prices. The comparative analysis and results clearly indicate that XGBoost outperforms the neural network model in terms of accuracy, with significantly lower error rates and computational time. Furthermore, the study reveals that Fare Class and the Market Share of the airline company are pivotal features affecting airfare prices. These insights enable airlines to refine their pricing strategies, particularly by focusing on increasing market share through adjustments in business class fares.

Interpreting these results, it's evident that XGBoost's robustness and efficiency make it a superior choice for airfare prediction compared to traditional neural network approaches. The findings address the research question of identifying the most effective predictive model for airfare prices, confirming that tree boosting methods, particularly XGBoost, offer a more accurate and efficient solution.

In terms of validity, the experiments conducted were comprehensive, covering various aspects of model performance, including error rates and feature importance. This thorough approach ensures the reliability of the conclusions drawn.

Looking forward, I plan to explore additional features that could further enhance the accuracy of the predictive model. One potential area of future work is the incorporation of real-time data, such as current economic indicators or sudden changes in airline policies, to see how they affect the model's predictions. Additionally, experimenting with more complex neural network architectures or other advanced machine learning techniques might offer insights into even more effective models for airfare prediction. The goal is to continually refine the model to adapt to the ever-changing dynamics of airline pricing strategies.

VIII. REFERENCES

[1] Timothy M. Vowles*, "Airfare Pricing Determinants in Hub-to-Hub Markets," Journal of Transport Geography 14 (2006) 15–22.

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

[3] 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.

[4] Benny Mantin*, Bonwoo Koo, "Weekend Effect in Airfare Pricing," Journal of Air Transport Management 16 (2010) 48–50.

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

[6] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," arXiv:submit/1502704 [cs.LG] 9 Mar 2016.

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

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

[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."