Optimizing Travel Costs: A Comprehensive Machine Learning Approach to Airfare Price Prediction

P.Kaveri[1], N.Naveen Kumar[2]
[1]Student MTech, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad
[2] Asst. Professor, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad

Abstract: The project delves into the intricate realm of global airline pricing strategies, leveraging advanced machine learning techniques to analyze and predict airfare prices across various airlines and destinations. By scrutinizing 136,917 flight records from prominent carriers such as Aegean, Austrian, Lufthansa, and Turkish Airlines, the study elucidates the multifaceted dynamics of pricing policies in the aviation industry. Utilizing a broad spectrum of machine learning models, including AdaBoost Regression, Gradient Boosting Regression, and Convolutional Neural Networks (CNNs), the research provides comprehensive insights into the complex interplay of factors influencing airfare pricing.

Despite challenges in data preprocessing, such as handling time attributes, the study showcases remarkable predictive accuracy, with certain models achieving up to 100% R2 score on specific datasets, notably the Austrian SKG-ARN route with Decision Tree Regression. These results underscore the efficacy of machine learning approaches in forecasting airfare prices, paving the way for enhanced consumer decision-making and industry optimization.

Index Terms: Airfare price, artificial intelligence, deep learning, machine learning, prediction model, pricing models, regression, quantum machine learning.

1. INTRODUCTION

The landscape of airline travel has undergone a dramatic transformation over the past five decades, transitioning from a luxury reserved for a select few to a ubiquitous mode of transportation accessible to a broad spectrum of travelers. Approximately 50 years ago, airline flights were primarily perceived as a luxury, with a focus on domestic routes rather than international ones. During this period, pricing policies for flight tickets remained static, reflecting a simpler era in the aviation industry's history.

However, as airlines sought to enhance profitability and expand their reach, they began to adopt management and economic software systems to optimize routes, adapt reservations, and introduce dynamic pricing strategies. One crucial advancement during this period was the implementation of yield management, a dynamic pricing strategy designed to optimize revenue through insights into, prediction of, and influence on consumer behavior. This marked a significant shift in the industry's approach to pricing, with airlines increasingly prioritizing customer preferences and experiences while simultaneously expanding their international destinations.

The convergence of market globalization and technological evolution has propelled airline companies into a highly dynamic and competitive environment, where traditional price optimization systems may struggle to keep pace with rapid changes. This necessitates the development of more sophisticated algorithms and software for dynamic price policy optimization. As a result, Artificial Intelligence (AI) algorithms have emerged as a promising solution for airfare price estimation, offering the potential for efficient and realistic results with faster processing speeds.

2. LITERATURE SURVEY

Airline pricing has been a subject of interest for researchers across various domains, from signal processing to artificial intelligence. This literature survey aims to provide an overview of the existing research on optimizing travel costs: a comprehensive machine learning approach to airfare price prediction, highlighting key contributions, methodologies, and insights.

K. Tziridis et al. (2017), In their paper titled "Airfare prices prediction using machine learning techniques," Tziridis and colleagues presented a study on predicting airfare prices using machine learning methods. They investigated the utilization of machine learning algorithms for predicting airfare prices, with a particular focus on the 25th European Signal Processing Conference (EUSIPCO) in 2017. Their work laid the foundation for subsequent studies in this area.

These studies collectively contribute to advancing the state-of-the-art in airfare price prediction, leveraging machine learning techniques to enhance accuracy and efficiency. By exploring diverse methodologies, datasets, and contexts, researchers have made significant strides towards developing robust predictive models that cater to the dynamic and complex nature of airline pricing dynamics.

However, challenges remain in addressing issues such as data sparsity, model interpretability, and scalability, highlighting avenues for future research in this domain.

3. METHODOLOGY

a) Proposed Work:

The proposed work aims to develop a comprehensive system for predicting airfare prices across six destinations served by four prominent airline companies. By leveraging a wide range of machine learning, deep learning, and quantum machine learning models, the system seeks to provide accurate predictions while offering insights into pricing policies and influential features. Drawing from a rich dataset of flight information, a holistic approach is employed to extract relevant features that capture the nuances of airfare dynamics.

From both destination and airline company perspectives, experiments are conducted to analyze the effectiveness of various algorithms, including AdaBoost Regression, Bagging Regression[28], Gradient Boosting Regression[29], Decision Tree Regression, Random Forest[30], ExtraTree[31], SVR[4], MLP[3], VGG11, VGG13, ResNet18, ResNet34, MobileNetV1, MobileNetV2[34], QSVR, and QMLP. By systematically evaluating these models, the proposed system aims to provide actionable insights for pricing strategies and optimize decision-making processes in the aviation industry.

b) System Architecture:

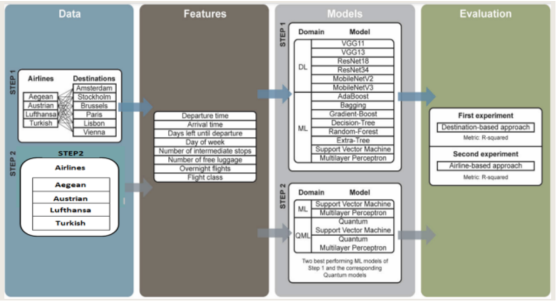


Fig-1: Proposed Architecture

The system architecture comprises two main steps: data preprocessing and model selection, followed by evaluation based on destination and airline-based approaches.

In the data preprocessing step, information is gathered regarding airlines (Aegean, Austrian, Lufthansa, Turkish) and destinations (Amsterdam, Stockholm, Brussels, Paris, Lisbon, Vienna). Features such as departure time, arrival time, days left until departure, day of the week, number of intermediate stops, number of free luggage, overnight flights, and flight class are extracted to build the dataset

For model selection, two sets of models are employed. In the first step, a comprehensive range of models from deep learning (VGG11, VGG13, ResNet18, ResNet34, MobileNetV2, MobileNetV3[35]) and machine learning (AdaBoost, Bagging[28], Gradient-Boost[29], Decision Tree[27], Random Forest[30], Extra-Tree[31], Support Vector Machine[4], Multilayer Perception) domains are considered.

In the second step, the focus shifts to machine learning models (Support Vector Machine, Multilayer Perception) and quantum machine learning models (Quantum SVM, Quantum Multilayer Perception).

Evaluation is conducted through two experiments: destination-based and airline-based approaches, with the metric being R-Squared. This approach allows for a comprehensive analysis of pricing policies and feature influences across different destinations and airlines, providing valuable insights for pricing optimization and decision-making in the aviation industry.

c) Dataset:

The dataset consists of flight information for six destinations served by four airline companies, organized into two experiments. In the first experiment, flights operated by Aegean, Austrian, Lufthansa, and Turkish airlines are recorded for each destination pair: SKG - AMS (Amsterdam), SKG - ARN (Stockholm), SKG - BRU (Brussels), SKG - CDG (Paris), SKG - LIS (Lisbon), and SKG - VIE (Vienna). This dataset includes features such as departure time, arrival time, days left until departure, day of the week, number of intermediate stops, and number of free luggage, overnight flights, and flight class.

In the second experiment, flights operated by the same four airlines are recorded without specifying the destination pairs. Instead, the dataset is structured based on the airline companies: Aegean, Austrian, Lufthansa, and Turkish. This dataset allows for an analysis of pricing policies and feature influences from an airline-based perspective.

Overall, the dataset provides a comprehensive view of airfare prices and related features across different destinations and airline companies, enabling analysis and prediction using machine learning, deep learning, and quantum machine learning models.

d) Data Processing:

Removing Duplicate Data: Duplicate data can skew analysis and model performance, so the first step is to identify and remove duplicates. By comparing records based on unique identifiers such as flight number or timestamp, duplicate entries can be identified and eliminated from the dataset, ensuring data integrity and accuracy.

Drop Cleaning: Drop cleaning involves removing or handling missing values, outliers, and irrelevant features to prepare the dataset for analysis. Missing values can be handled through imputation techniques such as mean, median, or mode replacement, or by dropping rows or columns with missing data. Outliers, which can distort statistical analysis, may be detected using methods like z-score or IQR (interquartile range) and either removed or transformed. Additionally, irrelevant features that do not contribute to the analysis or prediction task can be dropped to streamline the dataset.

By performing data processing steps such as removing duplicate data and drop cleaning, the dataset is refined and prepared for subsequent analysis and modeling. This ensures that the data used for training and evaluation is accurate, reliable, and representative of the underlying phenomena.

e) Visualization:

Seaborn and Matplotlib are powerful libraries in Python for creating visualizations that provide insights into data patterns and relationships. With Seaborn's high-level interface built on top of Matplotlib, complex visualizations can be created with ease.One common visualization technique is to create scatter plots to explore relationships between variables. For example, scatter plots can be used to visualize the relationship between departure time and airfare prices for different airline companies and destinations. Seaborn's `scatterplot()` function allows for easy customization of scatter plots with options to add color encoding for categorical variables or regression lines to visualize trends.

Another useful visualization technique is to create bar plots to compare categorical variables. For instance, bar plots can be used to compare average airfare prices across different days of the week or flight classes. Seaborn's `barplot()` function enables the creation of informative bar plots with options to customize colors, error bars, and confidence intervals.

Furthermore, line plots can be utilized to visualize trends over time, such as changes in airfare prices over different days or months. Seaborn's `lineplot()` function facilitates the creation of visually appealing line plots with options to add markers, smooth curves, and confidence intervals.Overall, Seaborn and Matplotlib provide a versatile toolkit for creating a wide range of visualizations that aid in data exploration, analysis, and interpretation.

f) Feature Engineering:

Feature Extraction: In the feature extraction process, the focus is on identifying and extracting relevant information from the dataset that can be used to train machine learning models effectively. One important step in this process is dropping time-based columns, as mentioned. While time-based features such as departure time and arrival time may contain valuable information, if they cannot be properly timestamped due to format discrepancies, it's prudent to exclude them to prevent negative values or inaccurate data.

Feature Selection: After extracting features, the next step is feature selection, where the most informative and relevant features are chosen for model training. This process helps in reducing dimensionality and computational complexity while improving model performance and interpretability. While time-based columns are dropped due to conversion issues, other features such as days left until departure, day of the week, number of intermediate stops, number of free luggage, overnight flights, and flight class can still be considered for analysis.

g) Training & Testing:

Dividing the dataset into training and testing subsets is essential for accurately evaluating the performance of machine learning models.

Training Data: The training dataset, representing approximately 70-80% of the total data, is used to train the machine learning models. This subset provides examples with known outcomes, allowing the models to learn the underlying patterns and relationships between the input features and the target variable. During training, the models adjust their parameters iteratively to minimize the error between their predictions and the actual values.

Testing Data: The testing dataset, comprising the remaining 20-30% of the data, serves as an independent set to evaluate the trained models' performance. This subset contains examples that the models have not been exposed to during training, simulating real-world scenarios. The models' predictions on the testing data are compared against the true labels to assess their accuracy, generalization ability, and robustness. This evaluation provides insights into how well the models can predict airfare prices for unseen data and helps identify any potential overfitting or underfitting issues.By splitting the dataset into training and testing subsets, we ensure that the machine learning models are trained and evaluated effectively, enabling reliable predictions and informed decision-making in the domain of airfare pricing.

h) Algorithms:

First Experiment

AdaBoost: AdaBoost is an ensemble learning technique that aggregates multiple weak learners to create a strong classifier. It iteratively adjusts the weights of misclassified instances to focus on difficult cases, thereby improving overall performance.

Bagging: Bagging, or Bootstrap Aggregating, Entails training several models on randomized subsets of the training data and merging their predictions to mitigate variance and enhance accuracy. It helps in creating more robust models by reducing overfitting.

Gradient Boosting: Gradient Boosting constructs a robust predictive model by sequentially incorporating weak learners, where each learner addresses the errors of its predecessors. It optimizes the loss function via gradient descent, leading to enhanced model performance.

Decision Tree: Decision Tree is a simple yet powerful algorithm that splits the data into subsets based on the value of input features.It recursively partitions the data into smaller subsets until a stopping criterion is met, producing a hierarchical structure employed for both classification and regression tasks.

Random Forest: Random Forest is an ensemble learning approach that creates numerous decision trees during training and outputs the mode of the classes (for classification) or the average prediction (for regression) from the individual trees. This method combats overfitting and boosts accuracy by combining predictions across multiple trees.

Extra Tree: Extra Trees, or Extremely Randomized Trees, is an extension of Random Forest where decision trees are built from random subsets of features and split points. It also diminishes variance through the introduction of randomness in the feature selection and splitting procedures.

SVR: Support Vector Regression (SVR) is a variant of Support Vector Machines (SVM) used for regression tasks.It finds the optimal hyperplane that maximizes the margin between the predicted values and the actual data points in a high-dimensional feature space.

MLP: Multilayer Perceptron (MLP) is a variant of feedforward neural network consisting of multiple layers of nodes (neurons) that can learn complex patterns in the data. It employs backpropagation to modifies the weights and biases during training, enabling it to approximate non-linear functions.

VGG11 and VGG13: VGG (Visual Geometry Group) is a deep convolutional neural network architecture known for its simplicity and uniform architecture. VGG11 and VGG13 refer to variants of VGG with 11 and 13 weight layers, respectively.
They consist of multiple convolutional layers subsequently max-pooling layers and fully connected layers. VGG networks are commonly used for image classification tasks due to their effectiveness in learning hierarchical features.

ResNet18 and ResNet34: ResNet (Residual Network) is an architecture of deep convolutional neural network designed to address the vanishing gradient problem in very deep networks. ResNet18 and ResNet34 are variants of ResNet[33] with 18 and 34 layers, respectively.
They introduce skip connections, or shortcuts, that allow the network to learn residual functions, facilitating train deeper networks. ResNet architectures are extensively utilised in image classification and object detection tasks.

MobileNetV1 and MobileNetV2: MobileNet is a family of lightweight convolutional neural network architectures designed for efficient computation on mobile and embedded devices.
MobileNetV1 and MobileNetV2 are two versions of MobileNet with different design strategies. MobileNetV1 uses depthwise separable convolutions to reduce the number of parameters and computational cost, while MobileNetV2 introduces inverted residual blocks with linear bottlenecks to improve performance. MobileNet architectures are commonly used for tasks where computational efficiency is critical, such as image classification on resource-constrained devices.
Combining these explanations gives a comprehensive overview of various deep convolutional neural network architectures commonly employed for image classification tasks, each possessing unique characteristics and advantages.

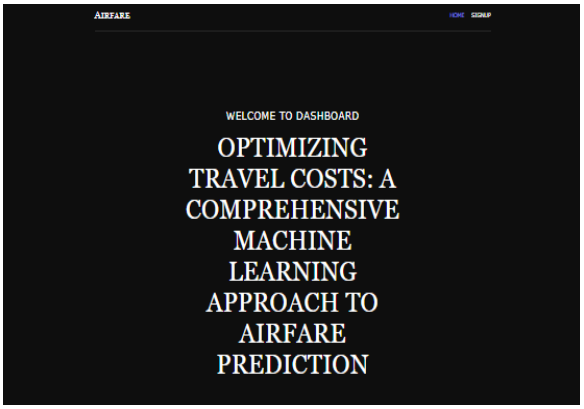
Second Experiment:
QSVR: Quantum Support Vector Regression (QSVR) is a quantum machine learning algorithm that extends classical SVR to quantum computers. It utilizes quantum computing principles to determine the optimal hyperplane for regression tasks, potentially offering advantages in computational efficiency and accuracy.

QMLP: Quantum Multilayer Perceptron (QMLP) is a quantum machine learning algorithm in light of the classical MLP architecture. It uses quantum circuits to represent and process data, enabling quantum parallelism and superposition to learn complex trends in the data.

4. EXPERIMENTAL RESULTS

R2 Score: R 2 = 1 − sum squared regression (SSR) total sum of squares (SST) , = 1 − ∑ (y i − y i ^) 2 ∑ (y i − y ‾) 2 . The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$



WELCOME TO DASHBOARD

OPTIMIZING TRAVEL COSTS: A COMPREHENSIVE MACHINE LEARNING APPROACH TO AIRFARE PREDICTION

Fig-2: Home Page

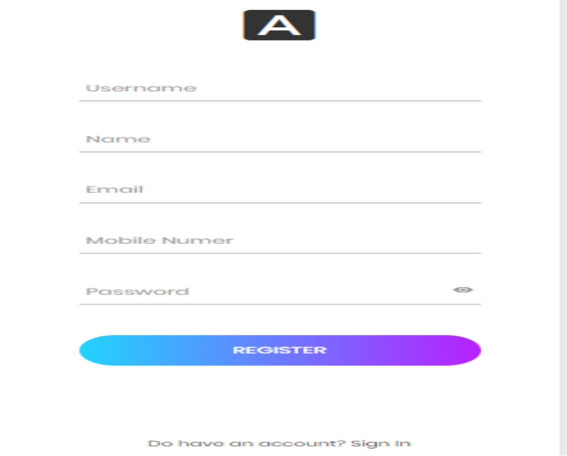
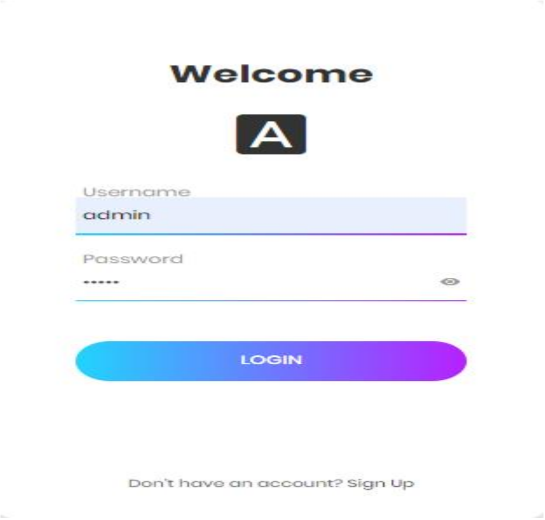


Fig-3: Registration Page

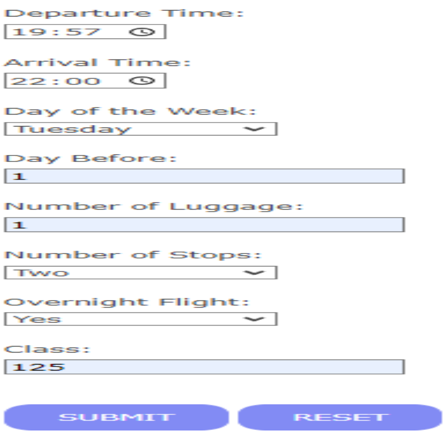


Welcome

Username
admin

Password
•••••

Fig-4: Login Page



Departure Time:
19:57

Arrival Time:
22:00

Day of the Week:
Tuesday

Day Before:
1

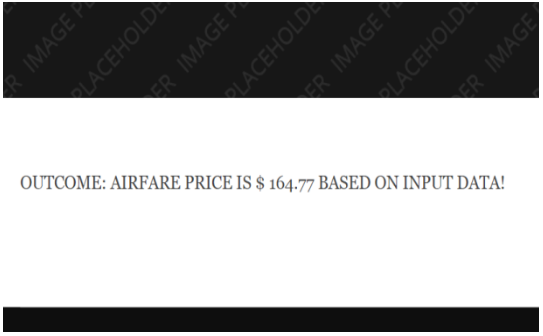
Number of Luggage:
1

Number of Stops:
Two

Overnight Flight:
Yes

Class:
125

Fig-5: Upload Input Data



OUTCOME: AIRFARE PRICE IS $ 164.77 BASED ON INPUT DATA!

Fig-6: Final Outcome

5. CONCLUSION

In conclusion, the project conducted a thorough evaluation of machine learning and quantum models for airfare price forecasting, providing insights into their predictive performance across diverse scenarios. Through comprehensive analyses encompassing various airlines, routes, and specific parameters, the project offered a detailed understanding of the strengths and weaknesses of different models.

The first experiment, which examined multiple airlines and routes, yielded nuanced insights into the predictive capabilities of various machine learning models. By utilizing R2 comparison graphs, the project provided a comprehensive assessment of each model's performance under varied conditions, contributing valuable knowledge to the field.

Furthermore, the second experiment focused exclusively on individual airlines, evaluating four distinct models—SVR, MLP, QSVR, and QMLP—for their effectiveness in predicting airfare prices. This focused analysis offered valuable insights into the performance of these models for specific carriers, guiding potential strategies for optimizing pricing decisions and enhancing overall accuracy.

Additionally, the project identified key parameters such as departure time, arrival time, and flight features that significantly impact airfare prediction accuracy. Understanding these factors is crucial for refining models and improving prediction performance in real-world applications. Moreover, the integration of Flask for user testing added practical value, providing a user-friendly interface for inputting parameters and visualizing airfare predictions, thereby enhancing the project's utility and usability.

Overall, the project's findings contribute to advancing airfare prediction methodologies, offering valuable insights for industry stakeholders and paving the way for more accurate and efficient pricing strategies in the aviation sector.

6. FUTURE SCOPE

Future research in airfare price prediction could explore additional features and datasets to further refine predictive models and enhance accuracy. Investigating customer segmentation based on flight features could provide valuable insights for targeted pricing strategies. Additionally, further exploration of quantum machine learning methods in airfare prediction, such as quantum Boltzmann machines, could offer efficient solutions for generating flight data and optimizing pricing strategies. Continued advancements in quantum hardware and computational resources are essential to unlocking the full potential of quantum models in real-world applications. Overall, the project lays the groundwork for future advancements in airfare price prediction, offering opportunities for more efficient pricing strategies and enhanced customer experiences in the airline industry.

REFERENCES

[1] S. Netessine and R. Shumsky, ''Introduction to the theory and practice of yield management,'' INFORMS Trans. Educ., vol. 3, no. 1, pp. 34–44, Sep. 2002.

[2] W. S. McCulloch and W. Pitts, ''A logical calculus of the ideas immanent in nervous activity,'' Bull. Math. Biophys., vol. 5, no. 4, pp. 115–133, Dec. 1943.

[3] F. Rosenblatt, ''The perceptron: A probabilistic model for information storage and organization in the brain,'' Psychol. Rev., vol. 65, no. 6, pp. 386–408, 1958.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik, ''A training algorithm for optimal margin classifiers,'' in Proc. 5th Annu. workshop Comput. Learn. theory, Jul. 1992, pp. 144–152.

[5] E. Fix and J. L. Hodges, ''Discriminatory analysis. Nonparametric discrimination: Consistency properties,'' Int. Stat. Rev./Revue Internationale de Statistique, vol. 57, no. 3, p. 238, Dec. 1989.

[6] R. E. Schapire, ''The strength of weak learnability,'' Mach. Learn., vol. 5, no. 2, pp. 197–227, Jun. 1990.

[7] K. Fukushima, ''Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,'' Biol. Cybern., vol. 36, no. 4, pp. 193–202, Apr. 1980.