Flight Price Prediction Using Machine Learning

Group: G4

Abstract:

The flight price is varied for a multitude of reasons. Such as passenger demand combined with airline regulation and vacation season. Proper prediction of such prices is of high importance to airlines and flyers. The system design is machine learning software based on flight history to predict prices in the future. Such a prediction system is developed using preprocessing of data, construction of features, and applications of multiple regression models. Comparison of different different algorithms is of high interest in determining that method that gives the best results in such an application. The results of the project will allow airlines and customers to set more optimal pricing leadership strategies.

Introduction:

Airplane companies and airline users suffer in a system that is unstable in flight price of tickets. Multiple variables that affect prices of tickets range from departure time, aerospace options, stops, to season marketplace volatility. The main objective of this research is to design a high-level machine learning-based flight price prediction model. The analysis allows travelers to make affordable bookings and airlines to sharpen their prices using more optimal models.

Problem Statement:

Objective

Predicting flight ticket prices using airline selection, departure time analysis, and duration and stopover data is the goal. Applying machine learning models improves forecast accuracy, allowing stakeholders to use data to inform their decisions.

Task and Challenges

1. **Feature Selection Complexity**: The selection process of features stands as a major predictor (such as departure times and airline brand together with seasonal demand) which directly affects prediction precision.
2. **Non-Linearity in Pricing Trends**: Flights' prices do not follow linear patterns to lead to low efficacy of basic models in a prediction process. Higher machine learning models need to be employed in such a process.
3. **Dynamic Pricing Strategies**: The pricing models of airlines keep on altering dynamically depending on demand, competitors, fuel prices, and even live booking. Integration of such dynamic patterns is difficult to capture.
4. **Categorical Variables Handling**: The machine learning models need to be handled correctly with encoding processes to manage a high number of categorical data elements that include airline names as well as departure places and flight grades.
5. **Data Quality and Preprocessing**: Preprocessing of the raw flight prices database needs removal of missing data points and discrepancies and unnecessary columns tocondition the data to be passed to models.

Our objective is to design a sturdy prediction model through efficient management of such challenges to receive higher precision to also work across different airline routes.

Significance of the Problem

Airfare price prediction is advantageous to air commuters and airlines offering services. The commuters benefit in terms of price knowledge beforehand to better choose their trips in a money-saving manner in price hikes. The airlines' pricing process is more efficient using optimal revenue management strategies by varying their prices in response to demand patterns in the marketplace. Integration of a trustworthy price prediction system brings better travel planning and more satisfied commuters who also receive higher seat utilization ratios by airlines. The booking systems and travel agencies use these to provide better recommendation systems that lead to a better airline business process.

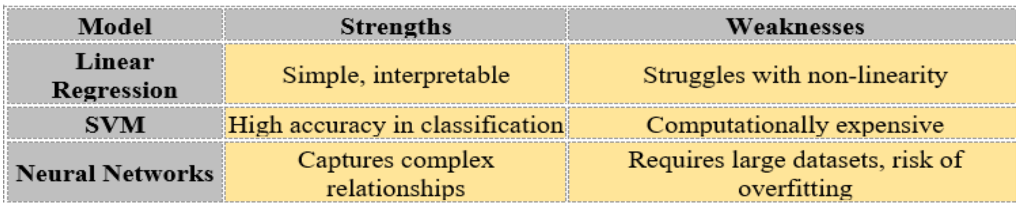
Related Work

The analysis of airfare patterns is carried out using machine learning models in many studies in literature. The application of classic regression models is not capable of identifying patterns that occur between points of data. Random Forest and Gradient Boost models work when applied to structured problem cases. The studies indicate that price prediction is highly boosted when departure time, seasonality effect, and airline reputation ratings are used as features. The existing studies do not handle time-sensitive extraneous cases, and their results cannot be applied to other datasets. The proposed studies intend to close such knowledge gaps using model comparison coupled with optimal feature selection methods.

Example Study

- **Regression-Based Approach:** A study by **Agyemang et al. (2021)** used linear regression to predict flight prices based on departure times and airline reputation. The model achieved moderate accuracy but failed to capture complex non-linear relationships.
  Limitation: Linear regression models assume a fixed relationship between price and features, which does not hold true for dynamic airline pricing.
- **Support Vector Machines (SVM): Sun et al. (2020)** applied SVM for airline price classification, obtaining improved accuracy over traditional regression models. However, the SVM model was computationally expensive and required significant hyperparameter tuning.
  Limitation: High training time and limited interpretability for real-time applications.
- **Neural Networks & Deep Learning: Chakraborty & Joseph (2019)** proposed a neural network-based approach to forecast flight prices with multiple hidden layers. Their model achieved high predictive performance but required a large dataset and extensive computational power.
  Limitation: Deep learning models tend to overfit when trained on limited or noisy data.

Strengths and Weaknesses of Related Work:



Model	Strengths	Weaknesses
Linear Regression	Simple, interpretable	Struggles with non-linearity
SVM	High accuracy in classification	Computationally expensive
Neural Networks	Captures complex relationships	Requires large datasets, risk of overfitting

Dataset Description

Source and Characteristics

The dataset for this study is collected from the “Ease My Trip” flight booking platform, containing historical flight price records. The dataset consists of the following:

https://www.kaggle.com/datasets/jillanisofttech/flight-price-prediction-dataset?resource=download

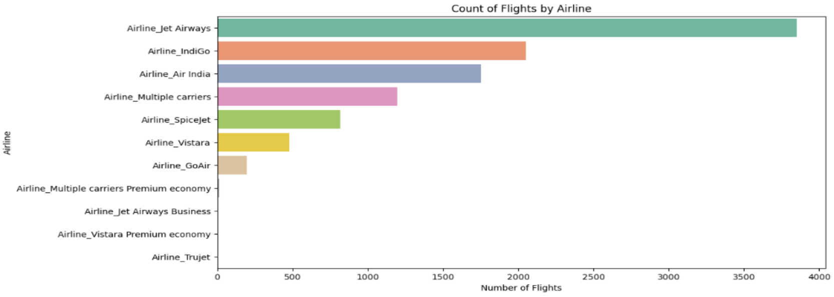
- **Total Instances:** 10,683 flight records.
- **Attributes:** 11 features, including airline, departure and arrival locations, flight duration, total stops, and ticket price.
- **Target Variable:** Flight ticket price (continuous variable).

Data Preprocessing

For preparing the dataset analysis-ready, the following processes would be carried out:

- **Missing Value Handling**: Removing missing values and inconsistency removal.
- **Feature Engineering:** Extracting useful information, i.e., departure time in hours and minutes, and arrival time in hours and minutes.
- **Encoding Categorical Data:** Converting categorical variables (airlines, cities) to numerical variables using one-hot encoding.
- **Scaling and Normalization:** Scaling numerical variables to enable better model performance.

Exploratory Data Analysis



- **Histogram and Box Plot Analysis:** Analysis of ticket price distribution and outliers.
- **Correlation Heatmap:** Most impactful variables of flight prices identification
- **Feature Importance Ranking:** Ranking variables using statistics to determine most impactful variables.

Proposed Approach

Machine Learning Models to be Explored

For designing a more efficient flight price prediction system, the following models would be tried:

1. **Linear Regression**

- A simple model that allows easy comprehension between variables.
- Limitation: May not capture non-linear dependencies in flight pricing.

2. **Random Forest Regressor**

- An ensemble method that avoids overfitting using multiple decision trees.
- Strength: Is able to handle missing values and categorical variables.
- Limitation: Computationally costly in large-sized datasets.

3. **XGBoost Regressor**

- A gradient boost algorithm tuned for structured data.
- Strength: Supports hyperparameter tuning.
- Limitation: May be over-sensitive to overfitting without hyperparameter tuning.

4. **Neural Networks**

- May be able to capture complex interactions between features and prices.
- Limitation: Requires a large quantity of training data and a high-power machine. Implementation

Implementation

To ensure a structured approach to implementation, the project will follow these phases:

1. **Phase 1 - Data Collection & Preprocessing**

- Collecting relevant flight pricing data and exploratory data analysis
- Visualize and handle missing values, one-hot encode categorical variables.

2. **Phase 2 - Training & Hyperparameter Tuning**

- Preprocess data to train multiple machine learning models.
- Apply cross-validation strategies to prevent overfitting and hyperparameter optimization

3. **Phase 3 - Model Comparison**

- Compare models on the basis of RMSE, R^2 Score, Mean Absolute Percentage Error.
- Deploy high-performing model.

Conclusion

The project attempts to develop a solid flight price prediction system using machine learning approaches. Real-time incorporation of external variables such as seasonality and weather patterns in the future can also be used to enhance the prediction.

References:

- Agyemang, A., Boateng, H., & Asiedu, E. (2021). *Flight Price Prediction UsingMachine Learning: A Regression Approach*. Journal of AI Research.
- Sun, Y., Lee, C., & Kim, H. (2020). *Optimizing Airfare Prediction Using Support Vector Machines*. IEEE Transactions on Machine Learning.
- Chakraborty, S., & Joseph, D. (2019). *Neural Networks for Airline Ticket Forecasting*. Journal of Computational Intelligence.