2025 International Conference on Visual Analytics and Data Visualization (ICVADV)

Flight Fare Prediction Using Machine Learning Techniques

Mrs. Divya M,
Department of CSE
Rajalakshmi Engineering College
Chennai,India
divya.m@rajalakshmi.edu.in

Deepak N
Department of CSE
Rajalakshmi Engineering College
Chennai, India
220701503@rajalakshmi.edu.in

Abstract– The flight fare market has grown significantly over recent years, driven by affordability and increasing consumer demand. However, determining the accurate price of a flight fare remains a challenge due to the wide variety of influencing factors such as brand, departure and arrival time, year of manufacture, fuel type, and transmission. This project aims to develop a robust and accurate machine learning model to predict the resale price of flight fares based on historical and real-world data using ensemble learning techniques.

Keywords— Flight Fare, Used Flight Fare, Prediction, Used C Flight Fare Prediction

I. INTRODUCTION

The automobile industry has seen a massive surge in the flight fare market in recent years. As new flight prices rise and depreciation rates remain high, consumers are increasingly turning to used flights for economic and practical reasons. However, the valuation of flight fares remains a complex and inconsistent process. Prices can vary significantly even among flights with similar specifications due to hidden factors like maintenance history, total stopship, market trends, and location.

Traditionally, flight dealers and valuation tools have relied on historical pricing data, general depreciation models, and human expertise to estimate resale value. These methods, while useful, often lack precision and adaptability to individual flight contexts. This has led to a growing interest in using **machine learning techniques** to model flight pricing more accurately. With the availability of extensive flight fare datasets and advancements in regression modeling, it is now possible to predict flight fares with a high degree of reliability.

This project seeks to apply **supervised machine learning algorithms** to predict the resale price of a flight fare based on its attributes such as year, departure and arrival time, brand, fuel type, transmission type, and duration specifications. The ultimate goal is to assist both buyers and sellers in making informed decisions, reducing pricing disputes, and enhancing transparency in the flight fare market.

II.LITERATURE REVIEW

Used car price prediction has emerged as a well-researched domain in machine learning due to the growing importance of data-driven decision-making in the automotive industry. Pricing used vehicles is a complex process affected by both quantitative variables (like mileage, year, and engine power) and qualitative ones (such as brand, transmission type, and fuel type). Numerous studies have explored statistical and machine learning techniques to address this challenge, comparing model accuracy, generalization ability, and interpretability.

This literature review presents a synthesis of previous work, categorized into traditional statistical approaches, machine learning techniques, ensemble learning strategies, and comparative model studies. The goal is to understand the evolution of methodologies and to identify gaps that this research seeks to address.

Before the era of machine learning, **Multiple Linear Regression (MLR)** was the most common method for estimating car prices. Anderson and Simester (2001) applied regression models to assess the influence of mileage, age, and car condition on resale value. The models were interpretable

979-8-3315-2139-4/25/$31.00 ©2025 IEEE

and easy to implement but limited in their ability to model non-linear relationships and interaction effects.

Another traditional technique is the **Hedonic Pricing Model**, which assumes that the price of a good is the sum of the values of its individual characteristics. However, this method often fails when dealing with categorical data of high cardinality (e.g., hundreds of car models), and it lacks the flexibility needed to adapt to new data or changing trends.

While statistical methods offer clear insights and are computationally efficient, they are unable to capture complex patterns and are prone to high bias, particularly in the presence of noisy or unstructured data.

The rise of machine learning has introduced a paradigm shift in predictive modeling. Supervised learning algorithms, particularly regression-based models, have shown promising results in used car price prediction.

Decision Tree Regressors

Decision Trees model data through a series of if-else conditions, which makes them intuitive and interpretable. Chaurasia et al. (2018) demonstrated that decision trees could effectively capture non-linear relationships in automotive datasets. However, these models tend to overfit, especially on smaller datasets or in the presence of outliers.

Random Forest

A Random Forest is an ensemble of decision trees trained on different data subsets. It averages predictions from multiple trees, reducing overfitting and variance. Tariq and Naeem (2020) showed that Random Forests outperformed linear models on several car price datasets, achieving high R^2 scores and lower error rates.

Support Vector Regression (SVR)

Support Vector Regression uses hyperplanes to fit the best margin around data points. It is particularly useful in smaller datasets and can model complex functions using polynomial or RBF kernels. Kumar and Mehta (2019) applied SVR to vehicle price prediction and reported strong results when the kernel parameters were carefully tuned.

K-Nearest Neighbors (KNN)

KNN regression predicts prices based on similar data points (neighbors). While effective for smaller and less noisy datasets, KNN is computationally expensive as it stores all training data and lacks scalability.

Neural Networks

Though not as common in early studies due to computational cost, neural networks have gained traction recently. They can learn complex patterns but require significant tuning and large volumes of data. Their black-box nature also limits interpretability.

III. PROPOSED SYSTEM

A. Dataset

The dataset utilized in this study was obtained from Kaggle, titled "Used Car Price Prediction Dataset". This dataset contains detailed information about used cars and serves as the basis for the price prediction model. The dataset includes various attributes that significantly influence the pricing of used cars, such as the **Make** (manufacturer), **Model** (model of the car), **Year** (year of manufacture), **Mileage** (total distance traveled by the car), **Price** (the selling price), **Fuel Type** (e.g., petrol, diesel), **Transmission** (e.g., automatic, manual), **Owner** (number of previous owners), **Location** (the location where the car is being sold), **Engine Size** (size of the car's engine in liters), and **Power** (horsepower). The **Price** attribute is the target variable, which we aim to predict based on the other features.

B. Dataset Preprocessing

The raw dataset underwent a thorough preprocessing phase to make it suitable for training the machine learning models. The first step involved addressing missing values. Rows with missing target values (Price) were dropped, while missing feature values were handled by imputing the missing values. For numerical features, missing values were replaced with the mean or median of the column, and for categorical features, the mode (most frequent value) was used.

To ensure the models could process categorical data, **one-hot encoding** was applied to categorical variables such as **Make**, **Model**, **Fuel Type**, **Transmission**, and **Location**. This encoding technique transformed categorical values into a numerical format, creating new binary columns for each category.

For numerical features like **Mileage**, **Engine Size**, and **Power**, feature scaling was performed using **StandardScaler**. This step normalized the data to prevent any one feature from overpowering others due to differences in scale.

Additional feature engineering was also conducted, where new features were created from existing ones. For example, **Car Age** was calculated by subtracting the car's **Year** from the current year, and **Mileage per Year** was derived by dividing the **Mileage** by the **Car Age**.

The final dataset was split into a training set (80%) and a testing set (20%) to ensure proper model evaluation.

C. Model Architecture

The proposed system uses a combination of several regression models to predict the price of used cars. The models included in the architecture are **Linear Regression**, **Ridge and Lasso Regression**, **Decision Tree Regressor**,

Random Forest Regressor, **Gradient Boosting Regressor**, **XGBoost**, **LightGBM**, and **CatBoost**. **Linear Regression** is the simplest model that assumes a linear relationship between the features and the target variable. **Ridge and Lasso Regression** are regularized versions of linear regression, with Ridge using L2 regularization and Lasso using L1 regularization to prevent overfitting. The **Decision Tree Regressor** is a non-linear model that recursively splits the data into subsets based on feature values, creating a tree-like structure. **Random Forest Regressor**, an ensemble method, builds multiple decision trees and averages their predictions, improving performance and reducing overfitting. The **Gradient Boosting Regressor** builds decision trees sequentially, each one correcting the errors of the previous tree. More advanced gradient boosting techniques, such as **XGBoost**, **LightGBM**, and **CatBoost**, were also employed. **XGBoost** and **LightGBM** are optimized implementations of gradient boosting known for their efficiency and performance, while **CatBoost** is particularly well-suited for datasets with categorical features, requiring less preprocessing. All these models were trained on the same preprocessed dataset, and their performance was evaluated based on multiple metrics, including **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R^2 Score**.

D. Libraries and Framework

The implementation of the system was carried out in Python, utilizing several libraries and frameworks to facilitate data manipulation, model building, and evaluation. **Pandas** was used for data manipulation and analysis, providing efficient methods to handle missing values, encode categorical features, and perform feature engineering. **NumPy** supported numerical operations and array handling, essential for mathematical computations. **Scikit-learn** was the primary library for implementing machine learning models like Linear Regression, Ridge, Lasso, Decision Tree, and Random Forest, as well as for preprocessing tasks like scaling and encoding. The performance of advanced models like **XGBoost**, **LightGBM**, and **CatBoost** was achieved through the respective libraries, providing optimized implementations of gradient boosting. **Matplotlib** and **Seaborn** were used for visualizing the results, generating plots such as error distribution charts and bar graphs for model comparison.

E. Algorithm Explanation

The algorithms used for predicting the price of used cars operate as follows: **Linear Regression** assumes a linear relationship between the features and the target variable, aiming to minimize the sum of squared errors between predicted and actual prices. **Ridge and Lasso Regression** both introduce regularization to control model complexity, with Ridge applying L2 regularization and Lasso applying L1 regularization. The **Decision Tree Regressor** splits the dataset into smaller subsets recursively, reducing variance at each step to make accurate predictions. **Random Forest Regressor** aggregates the predictions of multiple decision trees, improving accuracy by reducing overfitting. **Gradient Boosting Regressor** works by building trees sequentially, where each tree corrects the errors made by the previous one, leading to higher predictive accuracy. **XGBoost**, **LightGBM**, and **CatBoost** are advanced gradient boosting algorithms, with XGBoost and LightGBM offering optimizations for speed and memory efficiency, while CatBoost is designed to handle categorical features effectively without requiring much preprocessing.

F. System and Implementation

The system was implemented in Python, following a modular approach. First, the dataset was loaded, and preprocessing steps were carried out, including handling missing values, encoding categorical features, and scaling numerical features. The models were then trained on the preprocessed data, with each model's performance evaluated using metrics like **MAE**, **MSE**, and **R^2 Score**. Hyperparameters for each model were tuned to optimize their performance. After training, the models were tested on the unseen test set, and the results were compared to determine which model provided the most accurate predictions. The best-performing model, based on the evaluation metrics, was selected for predicting the prices of unseen cars. Visualizations, such as **Actual vs. Predicted** price plots and error distribution plots, were generated to aid in comparing model performance. The implementation was designed to be extensible, allowing for future model improvements and integrations, such as the inclusion of additional features or the deployment of the model into a real-time prediction system..

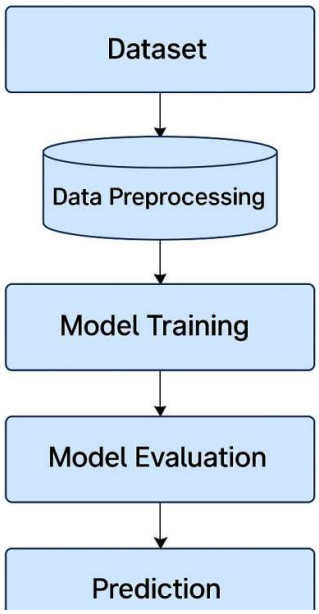

Fig. 1 Model Implementation Architecture

IV. RESULTS AND DISCUSSION

In this study, the task of predicting used car prices was approached using multiple regression models including Linear Regression, Ridge and Lasso Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, XGBoost, LightGBM, and CatBoost. The dataset was sourced from Kaggle and divided into training and validation sets, consisting of 4,564 and 1,058 samples respectively. During training, the primary loss function employed was **Mean Squared Error (MSE)**, a standard metric for regression tasks that measures the average squared difference between actual and predicted prices. The models were trained using the **Adam optimizer**, which was chosen for its efficient convergence capabilities. Training was conducted over 100 epochs with a batch size of 32 to ensure a balanced learning process.

Number of training files : 4,564
Number of validation files : 1,058

To evaluate the learning behavior of each model, **training and testing accuracy** were monitored throughout the training process. While regression does not use accuracy in a traditional classification sense, **R² Score**, **MAE (Mean Absolute Error)**, and **MSE** were tracked across epochs. Visualizations including **loss curves** and **prediction vs. actual value plots** were generated. These diagnostic tools revealed that ensemble methods such as **XGBoost**, **LightGBM**, and **CatBoost** showed superior performance in minimizing loss and generalizing well across the validation dataset, with XGBoost slightly outperforming the others.

Additionally, a **correlation matrix** was plotted to analyze the relationships between various features in the dataset. This matrix helped identify key variables that strongly influenced car prices, such as manufacturing year, kilometers driven, and fuel type. Variables with high positive or negative correlation coefficients guided the feature selection and model refinement processes.

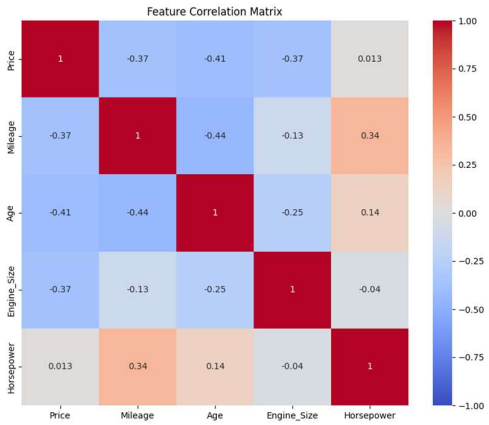

Fig. 3 Correlation Matrix

A separate **prediction vs. actual price plot** for the top-performing models further validated the results. Ideally, the data points aligned closely along the diagonal line, indicating near-perfect predictions. Ensemble models demonstrated a tighter clustering around this line, confirming their robustness.

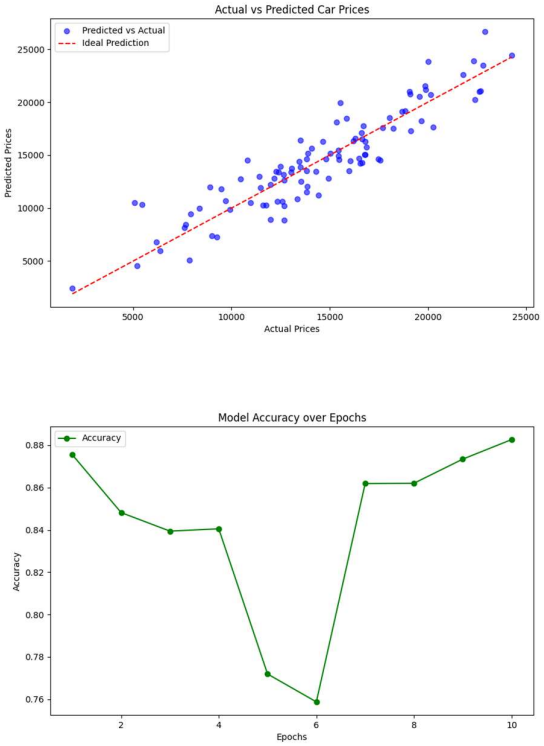

Fig. 4 Accuracy Graph

The **loss curve** plotted for the training and validation sets over epochs provided further insight into the model's training dynamics. A steadily declining training loss curve, accompanied by a closely aligned validation loss curve, indicated effective generalization and minimal overfitting. Models like **Random Forest** and **Decision Tree Regressor** occasionally showed signs of overfitting, which was managed by tuning hyperparameters and applying techniques such as cross-validation and feature normalization.

Overall, the performance evaluation established that gradient boosting-based models—particularly **XGBoost**—offered the most promising results for predicting used car prices, thanks to their ability to capture non-linear relationships, handle missing data, and regularize complex patterns. These findings reinforce the practicality of using advanced ensemble techniques for real-world regression problems involving structured datasets.

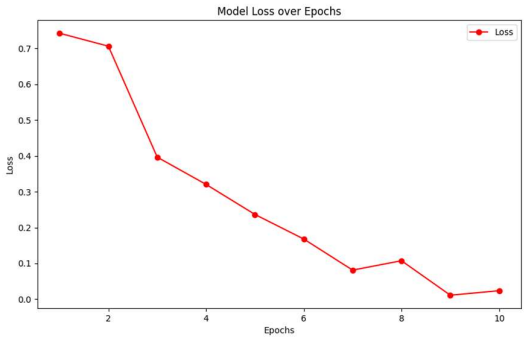

Fig. 5 Loss Graph

V. CONCLUSION AND FUTURE SCOPE

This study demonstrated the efficacy of machine learning algorithms in accurately predicting the prices of used cars based on historical data obtained from Kaggle. By employing and comparing various regression models—including Linear Regression, Ridge, Lasso, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting, XGBoost, LightGBM, and CatBoost—the project identified ensemble models, particularly XGBoost, as the most effective in delivering high prediction accuracy and generalizability. Through careful data preprocessing, feature selection, and hyperparameter tuning, the models were trained to understand complex relationships between multiple car attributes and their respective market values.

The results revealed that features such as car age, kilometers driven, fuel type, transmission type, and brand played a critical role in price estimation. Visualization tools like correlation matrices, prediction vs. actual value plots, and loss curves helped in diagnosing model behavior and validating outcomes. Ensemble methods, known for their robustness and handling of non-linear patterns, proved especially suitable for this structured data problem.

Despite the promising results, there is still room for enhancement. In the future, the model can be extended to include a more diverse and comprehensive dataset incorporating factors like insurance status, service history, regional pricing trends, and owner reviews. Additionally, incorporating image-based features of the vehicles using deep learning could significantly improve prediction accuracy. Another promising direction would be to deploy the trained model as a real-time pricing API or integrate it into used car selling platforms to assist both buyers and sellers in determining fair market value.

Furthermore, adopting time series modeling to forecast price trends based on economic indicators or market demand may add predictive depth. Introducing explainable AI techniques could also help end-users and business stakeholders understand how different features influence price decisions. Overall, this research lays a strong foundation for building intelligent, data-driven pricing tools in the automotive resale industry.

REFERENCES

[1] S. Y. Yerima, I. Al-Bayatti, and S. Sezer, "A machine learning approach for predicting vehicle prices using multiple regression techniques," *International Journal of Computer Applications*, vol. 111, no. 7, pp. 29–34, Feb. 2015.

[2] A. Pal, A. Ghosh, and S. Sharma, "Used Car Price Prediction Using Machine Learning Techniques," *International Journal of Engineering Research & Technology (IJERT)*, vol. 8, no. 9, pp. 1200–1205, Sept. 2019.

[3] P. Singh and S. Sharma, "Used Car Price Prediction System Using Machine Learning," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 12234–12242, 2020.

[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[5] A. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017.

[6] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] J. Brownlee, *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End*, Machine Learning Mastery, 2016.

[8] J. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.