Used Car Price Prediction Project Report

1. Introduction

The used car market is vast, with prices influenced by numerous factors like brand, mileage, model year, and engine specifications. This project aims to create an efficient machine learning model that predicts used car prices, providing valuable insights for buyers, sellers, and market analysts. The dataset for this project was sourced from Kaggle and includes features such as brand, model, engine specifications, mileage, and accident history.

2. Data Exploration and Preparation

The dataset contained information on used cars, including features like brand, model, engine details, model year, mileage, accident history, and transmission type. During data exploration, key features and patterns were identified, such as the relationship between model year and price and the impact of engine horsepower on the overall price.

To prepare the data for training, the following steps were taken:

- **Horsepower Extraction:** Extracted horsepower information from the engine column.

- **Feature Encoding:** Categorical variables were encoded using TargetEncoder for better representation in the models.

- **Data Scaling:** StandardScaler was used to standardize numerical features.

- **Feature Engineering:** Derived features such as the age of the vehicle ('vehicle_age') were introduced to enhance model performance.

3. Model Training

Three different machine learning models were trained to predict used car prices:

- **CatBoost Regressor**: A gradient boosting framework that handled categorical features well. This model achieved an RMSE of **67,838.64** on the validation dataset.

- **LightGBM Regressor**: A gradient boosting algorithm that focused on faster training. With hyperparameter tuning using `RandomizedSearchCV`, the model achieved an RMSE of **67,549.40**.

- **Ensemble Approach**: Models were ensembled to leverage their strengths. LightGBM performed slightly better and was selected as the final model.

4. Evaluation and Results

The model evaluation was based on the **Root Mean Squared Error (RMSE)**. The results of each model are as follows:

- **CatBoost Regressor**: RMSE = **67,838.64**

- **LightGBM Regressor**: RMSE = **67,549.40**

- **Stacked Ensemble**: The stacked ensemble approach did not outperform LightGBM due to overfitting, leading to an RMSE of **68,652.35**.

The LightGBM model provided the best performance, making it the final choice for generating price predictions.

5. Model Deployment

The final LightGBM model was used to predict prices for unseen test data, and the results were saved to a CSV file (`test_predictions.csv`). These predictions can assist buyers in identifying fair prices and sellers in setting competitive prices.

6. Challenges and Future Improvements

- **Handling Large File Sizes**: During the project, large file sizes posed challenges in version control. Utilizing Git LFS (Large File Storage) can help overcome such issues in future projects.

- **Hyperparameter Optimization**: Bayesian optimization could further enhance model performance, though there were difficulties in setup. Implementing this in future iterations might yield better results.

- **Additional Features**: Adding more features like previous ownership details or service history could improve the model's accuracy.

7. Conclusion

The project successfully developed a model to predict used car prices with reasonable accuracy. Using a combination of data preprocessing, feature engineering, and multiple machine learning models, the best-performing model (LightGBM) achieved an RMSE of **67,549.40**. This project provides a framework for price prediction in the used car market, which can be extended with additional features and more advanced optimization techniques.

8. References

- Kaggle dataset: [Playground Series - Season 4, Episode 9](https://www.kaggle.com/datasets)

- CatBoost documentation: https://catboost.ai/

- LightGBM documentation: https://lightgbm.readthedocs.io/