Lambton
College

USED CAR-PRICE PREDICTION

USING MACHINE LEARNING

PROJECT PROPOSAL

BDM 3014
Introduction to Artificial Intelligence

SUBMITTED TO:
Bhavik Gandhi
Lambton College

MARCH, 2024

SUBMITTED BY
Group: 2
Big Data Analytics (Second Sem)
Lambton College

I. INTRODUCTION

In the rapidly expanding and competitive landscape of the global used car market, where approximately 40 million vehicles are sold annually, effective pricing strategies play a pivotal role in ensuring efficient sales and profitability. For instance, in India, the market, predicted to grow at a CAGR of 15.12% from 2020 to 2025, demands innovative solutions to address the dynamic pricing challenges. This project endeavours to revolutionize the used car market by leveraging AI and ML techniques to enhance the precision of price predictions.

The project delves into the complexities of the automotive sector in the world, where the requirement for used automobiles is increasing. Existing online platforms like OLX and Quickr have influenced the market, potentially distorting pricing dynamics. To counteract these distortions, the project explores the application of advanced algorithms to forecast accurate automobile values based on specific criteria. As the automotive industry undergoes shifts, marked by a notable decrease in new vehicle production, there is a discernible preference for used cars, emphasizing the need for a standardized pricing mechanism.

Multiple Machine Learning (ML) algorithms are proposed to predict used vehicle values, drawing on historical data and aggregating prices from diverse sources. The dataset used in this experiment encompasses crucial factors such as the make year, model, fuel applicable, transmission, mileage, and ownership history. By training the model on this comprehensive dataset, the project aims to provide reliable and precise predictions, bridging the gap between sellers and buyers in the market.

In response to the growing demand for second-hand vehicles, this project aims to develop an AI solution that prioritizes customer-friendly pricing. By analyzing data from various vendors and purchasers, the machine learning model seeks to establish fair and accurate pricing for used automobiles. The goal is to contribute to the standardization of the used car market around the globe, ensuring transparency and reliability in pricing through cutting-edge technology.

II. PROBLEM STATEMENT

Let’s consider purchasing a second-hand automobile. When looking for a certain model and year, you may come across an advertisement for the ideal vehicle. You're unclear if the asking price is realistic. Comparing the costs of identical automobiles may be time-consuming and inefficient. We propose software that uses basic car information to produce an acceptable asking price based on prior sales.

This project proposes a method that allows users to input basic information about a specific vehicle into a program that generates a reasonable asking price based on previous sales. The project's target consumers are those who purchase or sell used cars. A more open and efficient pricing structure will help both buyers and sellers of used cars. This would allow buyers to determine if an asking price is above average, below average, or fair, while sellers can easily determine a reasonable asking price for their vehicles. The project addresses the challenge of identifying the most important factors in the price prediction model, which helps both parties better understand the reasoning behind the current value of their vehicles.

III. PROJECT OUTCOMES

- **Informed Buying and Selling:** Helps consumers make informed decisions when buying or selling used cars.
- **Price Transparency:** Promotes transparency by showing how various factors affect car prices.
- **Fair Pricing:** Facilitates fair pricing in the used car market for both buyers and sellers.
- **Time and Effort Savings:** Saves time and effort by automating the price estimation process.
- **Reduced Information Asymmetry:** Reduces information gaps between buyers and sellers.
- **Market Insights:** Provides insights into market trends for dealerships and the automotive industry.
- **Enhanced User Experience:** Offers a user-friendly tool for estimating car prices.
- **User Feedback and Improvement:** Uses user feedback to improve price predictions over time.

IV. DIFFERENTIATION

The **Used Car Price Prediction Model** will stand out from other solutions available in the market due to its comprehensive approach to predicting used car prices. Unlike other models that might focus on a limited set of features, this project will consider a wide range of factors that influence used car prices. Moreover, the model will be built using advanced machine-learning techniques like Random Forest Regression. This high accuracy and the inclusion of a broad set of features make the model a valuable tool for both sellers and buyers in this market. Our model offers several unique advantages over other solutions:

- **Customized Model:** Our model can be trained specifically on the data that is most relevant to our users, such as local market data, specific car brands, or types of cars.
- **Continuous Learning:** Our model can improve over time. As more data becomes available, our model can be retrained to reflect the latest trends and patterns in car pricing.
- **Transparency:** By building our model, we have full transparency into the factors that the model considers important in predicting car prices.
- **Integration:** Our model can be seamlessly integrated into our existing systems or applications.
- **Cost-Effective:** While there might be costs associated with developing and maintaining our model, it can be more cost-effective in the long run compared to paying for third-party solutions.

IV. PROJECT COMPONENTS

This project contains three-phase, as elaborated below:

1. Data Collection/Gathering Phase: This module is responsible for gathering data about used cars. The dataset for this project will be obtained via Kaggle, and scraped from Craigslist, the world’s largest collection of used autos for sale. The dataset consists of 423,857 rows and 25 features, including a continuous dependent variable (“price”) that we aim to predict. The features used in the dataset are:

- Registration Year
- Selling Price
- Model
- KMs Driven
- Fuel Type
- Location
- Transmission
- Mileage

2. Model Building Phase: After collecting the data, the data is preprocessed. Based on this a machine learning model is tested with different models with different hyperparameters and select the best model. We will start the phase following the complete life cycle of data science, including all the steps:

a) Data Preprocessing: This is the preliminary step in any machine learning project. Raw data collected from various sources often requires cleaning and preprocessing. This module cleans the data by handling missing values, outliers, and inconsistencies. It may also encode categorical variables and normalize or scale numerical features.

- **Removing outliers:** Outliers can significantly skew the model’s understanding of the data, leading to inaccurate predictions. We may deploy the techniques same as Z-score or IQR method to identify and remove these.
- **Handling missing values:** Values that are missing may be addressed using a variety of ways, including filling them with the column's mean, mode, and median or using a predictive filling algorithm. In cases where the values are not needed, the rows or columns with missing values may be dropped.
- **Handling duplicates:** Duplicate entries can bias the model towards certain observations. Identifying and removing these duplicates is crucial to ensure the model’s accuracy.

b) Feature Engineering: In this module, relevant features or variables are selected and transformed to improve the model's ability to predict car prices. For example, you might create new features like "age of the car" or "average market price of cars of the same make and model."

- **Encoding categorical variables:** Machine learning models require numerical data. Categorical variables can be converted to numerical data using methods such as one-hot encoding or ordinal encoding.

- **Scaling numerical features:** Certain ML algorithms may encounter problems when dealing with features of varying sizes. Scaling uses procedures like standardization and normalization to bring all characteristics to the same scale.
- **Handling highly correlated features:** Highly correlated features can lead to overfitting. Techniques such as variance inflation factor (VIF) can be used to identify and remove these.

c) **Model Training:** This module involves the development and training of machine learning models for price prediction. Common models used for regression tasks in this project include linear regression, decision trees, random forests, support vector regression, or more advanced techniques like gradient boosting and neural networks. Multiple models may be trained and evaluated to select the best-performing one. This step involves:

- **Splitting the dataset:** Typically, the dataset is divided into two parts: training and testing. The training set trains the model, while the test set evaluates its performance.
- **Feature scaling:** Some machine learning methods demand that the features be on the same scale. Methods such as standardization and normalization can be applied to this.
- **Hyperparameter** This entails changing the settings of the machine learning algorithm to increase its performance. Random search and Grid search techniques are also viable options.

d) **Model Evaluation and Validation:** This module verifies the performance of the machine learning models to guarantee that they are accurate. It evaluates the model's ability to forecast automobile prices using cross-validation techniques and measures such as mean absolute error (MAE) or mean squared error (MSE). This includes:

- **Using metrics:** Metrics like R-squared and accuracy may be used to evaluate the model's performance. The measure used depends on the situation at hand.
- **Achieving a high accuracy rate:** The objective is to create a model that can reliably anticipate automobile prices. This might entail going back and revising the preprocessing, feature engineering, or training processes.
- **Prediction:** This entails utilizing the trained and evaluated model to generate predictions about new data. The model takes in a car's attributes and generates a forecasted price.

e) **User Interface (UI) Module:** This module provides the user interface for data input and price estimation. It's typically implemented as a web application or mobile app. Users enter details about the used car, and the estimated price is displayed. The module communicates with the backend for processing.

f) **Backend Processing Module:** The backend processes user input applies the machine learning model to generate a price estimate and returns the result to the user interface. It may also include functions for data retrieval, model loading, and result presentation.

g) **User Feedback Module:** This optional module gathers user feedback and ratings on the accuracy of the price estimates. The feedback can be used for model refinement and improvements over time.

h) Deployment Module: Once the model is trained and the system is developed, it needs to be deployed to a server or cloud platform, making it accessible to users.

i) Database Module: A database is used to store historical car data, pricing models, and user feedback. It helps in data retrieval and model updates.

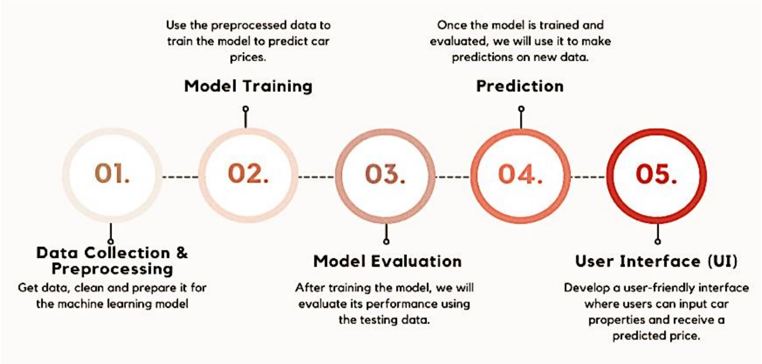
Use the preprocessed data to train the model to predict car prices.
Model Training
Once the model is trained and evaluated, we will use it to make predictions on new data.
Prediction
01.
02.
03.
04.
05.
Data Collection & Preprocessing
Get data, clean and prepare it for the machine learning model
Model Evaluation
After training the model, we will evaluate its performance using the testing data.
User Interface (UI)
Develop a user-friendly interface where users can input car properties and receive a predicted price.

Fig: Flowchart of Analytical/Mathematical modules

3. Other important Components:

a) Documentation: Comprehensive documentation will be provided to explain the model, its features, and how to use it.

b) Deployment: The model will be deployed on a cloud platform, ensuring users can access it at any time.

c) Maintenance and Updates: Regular maintenance ensures the model is accurate and relevant.

VI. PROJECT REQUIREMENTS

- **Hardware requirements:**
  Processor: Intel(R) Celeron(R) CPU N3050 with Clock Speed: 1.60 GHz
  RAM Capacity: 3.92 GB
  Operating System: 64-bit
  Processor Architecture: x64-based

- **Software requirements:**
  Python: One of the most used programming languages

- **Overview of Programming Tools/Libraries:**
  - **Notebook:** a free, open-source, interactive online application for creating Python code.
  - **NumPy:** an open-source numerical Python library that supports multidimensional arrays and matrix data structures.
  - **Pandas:** a widely used data science and machine learning tool built on NumPy.
  - **Matplotlib:** a Python library for plotting graphs using NumPy and Pandas.
  - **Seaborn:** a Python library for plotting graphs using Matplotlib, Pandas, and NumPy.
  - **Skeikit-learn:** a powerful tool for machine learning and statistical modelling.

- **Scipy.stats:** a module containing probability distributions and a growing library of statistical functions.

VII. CONCLUSION

To conclude, the creation of the Used Car Price Prediction Model represents a big leap in tackling the complexities of the dynamic used car market. By incorporating a broad spectrum of features and employing sophisticated machine learning techniques, this project provides a valuable tool for both sellers and buyers in need of accurate pricing information.

The model's distinctive advantages stem from its adaptability to local market data, specific car brands, or types of vehicles, offering users customized insights. Its ability to learn continuously ensures it remains responsive to changing market trends, facilitating regular updates, and enhancing accuracy over time. The transparency inherent in the model's design empowers users with a clear comprehension of the factors influencing price predictions.

Moreover, the model's seamless integration into existing systems or applications enhances user accessibility, while its cost-effectiveness, in comparison to third-party solutions, highlights its long-term value. The dataset's sourcing from Craigslist, combined with thorough data preprocessing, ensures a solid foundation for training the model. The project's numerous components, from data preparation and feature engineering to model training, assessment, and deployment, show a methodical approach to creating a strong solution. The inclusion of a user-friendly interface, along with comprehensive documentation, enhances the model's usability and encourages educated decision-making in the used automobile market.

Concerning industry growth predictions and prior studies, the model emerges as a pioneering effort to navigate the intricacies of pricing used cars, offering a reliable and innovative tool for stakeholders in the automotive industry. The Used Car Price Prediction Model will provide a valuable tool for both sellers and purchasers in the second-hand car market, helping them make informed decisions about the pricing value of such cars. By leveraging advanced ML techniques and a comprehensive set of features, this project intends to overcome the issues associated with pricing used automobiles and create a market-leading solution.

REFERENCES

- Listiani, M. (2009). Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Thesis (MSc). Hamburg University of Technology.
- Oprea, C. (2010). Making the decision on buying second-hand car market using data mining techniques (Special), pp.17-26.
- Ozgur, C., Hughes, Z., Rogers, G., & Parveen, S. (2016). Multiple Linear Regression Applications in Automobile Pricing. International Journal of Mathematics and Statistics Invention, pp.01-10.
- Lessmann, S., Listiani, M., & Voß, S. (2010). Decision support in car leasing: A forecasting model for residual value estimation.
- Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car Price Prediction Using Machine Learning. TEM Journal, 8(3), 830–836.
- Sun, N., Bai, H., Geng, Y., & Shi, H. (n.d.). Price Evaluation Model In Second-Hand Car System Based On BP Neural Network Theory. Hohai University Changzhou, China.
- Monburinon, N., Rungpheung, S., Buya, S., & Boonpou, P. (2018). Prediction of Prices for Used Car by using Regression Models. In ICBIR 2018.