**CS 57600-01 Machine Learning**

Predicting Car Prices using Machine Learning

Team Members: Aakriti Deora ([deora01@pfw.edu](mailto:deora01@pfw.edu))

Rudraksh Sugandhi ([sugar01@pfw.edu](mailto:sugar01@pfw.edu))

1. **Introduction** :

Machine Learning and Data science has brought very crucial positive shift in the current technology market. There are various applications which are ranging from self-driving cars, predicting manufacturing defects, healthcare, identification of news as real or fake and many other. Predicting car prices using Machine Learning is one of its applications which aims at estimating the cars prices. Anyone can use this technology to determine the price of the car more accurately based on the market value of the cars. This will result in setting the price of the car accurately from seller’s and buyer’s perspective. COVID-19 has disrupted the production of chips used in car manufacturing, which resulted in car prices fluctuating heavily in both new and resale markets. Based on the existing data, this project’s aim is to use the machine learning algorithm to develop a model for predicting the car prices. The model will extract interesting insights and make best predictions.

The following sections introduces about Problem Statement and related work that has been carried out in this field which are documented are also discussed in the following sections. Also, some background knowledge with respect to the steps carried out along with the data description, pre-processing techniques are described.

1. **Problem Statement** :

To develop a model that predicts the price of the car accurately. The algorithm used here is Random Forest regression. As Random Forest regression is a supervised learning algorithm that uses ensemble learning method for learning. Ensemble learning method is a technique that combines predictions from multiple learning algorithms to make an accurate prediction than a single model. Data set contains features like Name\_of\_car, Year\_of\_purchase, selling\_price, miles\_driven, fuel\_type, seller\_type, transmission: Manual or automatic, Owner\_info. In future other algorithms can be implemented and results will be compared based on the prediction accuracy.

1. **Related Work**:
   1. *AI Blue Book: Vehicle Price Prediction using Visual Features*

**Problem Description**: In this work, The Ai Blue book team is creating a series of machine learning models that estimate the price of a product based on its image, as well as show the features that result in greater or lower price predictions.

**Application:** We have found several potential real-world uses for this work. Using feature visualization, merchants can determine which qualities of objects are associated with higher costs, and then utilize this information to help advise and steer product creation. These models can also be expanded to help with large-scale valuations, such as used car sales, where a large number of products must be evaluated fast, and pricing might be difficult to calculate.

**Machine learning Methodology:** The Paper described the model developed baselines for price regression using linear regression on histogram of oriented gradients (HOG) and convolutional neural network (CNN) features, as well as a baseline for price segment classification using a multiclass SVM. For main models, they train several deep CNNs for regression and classification using both transfer learning and their own architectures. On both datasets, deep CNNs outperform other models in a variety of metrics. They create their own deep learning architecture, PriceNet, which is based on the Squeeze Net architecture. Because expensive 3x3 convolutions are replaced with fire modules, SqueezeNet has fewer parameters than other model architectures. The depth of the volume is first down sampled by efficient 1x1 convolutions (squeeze), then expanded by a combination of 1x1 and 3x3 convolutions (expand) in a fire module. The SqueezeNet architecture is then modified by adding residual connections between the fire layers and batch normalization to each fire module. There are approximately 1.2 million parameters in networks. PriceNet Architecture:

Chart, bar chart

Description automatically generated

**Experimental data:** The cars dataset contains images of vehicles as well as their prices. They retrieve price data from Kaggle and then combine these prices with images from Google Images, using search terms that include model and year, as well as "Angular Front View." The images are then cleaned and resized, yielding a final dataset of 1,400 examples.

**Evaluation methods:** To ensure fair comparison, they split the datasets into training and testing splits that are consistent across all models. To create the split for both datasets, shuffle the points first, then assign 90 percent to train and the remaining 10% to test. Then, using three different metrics, we evaluate and compare the price regression performance of our models: root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R2).

**Results:** PriceNet, custom network architecture, outperforms multiple transfer learning and linear regression baselines in regression. The transfer learning deep network outperforms models for price category classification significantly. Moreover, they utilize three different methods to visualize which image regions deep CNN models discriminate price with, providing insight into which visual features of products result in certain prices.

* 1. *Price Prediction of Used Cars Using Machine Learning*

**Problem Description:** This paper aims to develop a model that predicts reasonable used car prices based on a variety of factors such as vehicle mileage, year of manufacture, fuel consumption, transmission, road tax, fuel type, and engine size.

**Applications:** In the used car market, this model can benefit sellers, buyers, and car manufacturers. Upon completion, it can generate a reasonably accurate price prediction based on the data entered by users. Machine learning and data science are used in the model-building process.

**Machine learning Methodology:** This examination of information several regressions to see which one produced the best results, specifically which regression fit the dataset the best. To assess the model's performance, the R-square was calculated for each regression. R-square, also known as the coefficient of determination, is a common statistic. The regression with the highest R-square score was chosen after calculating each regression's R-square.

1) Linear Regression

2) Polynomial Regression

3) Support Vector Regression (SVR)

4) Decision Tree Regression

5) Random Forest Regression

**Experimental data:** The dataset used was scraped data from 100,000 used cars in the United Kingdom from Kaggle user Aditya. Each row represented a specific used vehicle on the market. Each column represented a different aspect of the vehicle, such as the model, year, selling price, transmission, mileage, fuel type, tax, mpg, and engine size. Audi, BMW, Ford, Hyundai, Mercedes, Toyota, Vauxhall, and Volkswagen vehicle data were separated by car manufacturer. This project made use of the Mercedes component, which contained 13,120 rows of Mercedes vehicle information. This project intended to take into account all nine factors when building the network.

**Evaluation methods:** The R-square was calculated for each regression to evaluate the model's performance.

**Results:** The R-square for linear regression is 0.72354, which is the lowest of the five models. As a special case or improved version of linear regression, polynomial regression performs significantly better than linear regression. It has an R-square of 0.83127, which is an increase of more than 10%. Decision tree regression achieves a 0.85140 score, which is marginally better than SVR. The square increased to 0.90416 when we switched to random forest regression. Compared to the other four regressions, it is significantly higher. This indicates that the random forest regression model best fits the data structure and provides the most accurate predictions.

* 1. *Second Sale Car Price Prediction using Machine Learning Algorithm*

**Problem Description:** In this paper, a prediction model is built so that it estimates the selling price of already used cars based on their features. This in turn reduces the burden and risk from the seller, consumer and also provides a positive first - hand knowledge about price and low financing cost for used cars.

**Applications**: Many big organizations are majorly handling the present market for used car sales.

Some of the major features to consider while determining the price of used cars include the brand name, year, kilometers driven, owner and seller type, fuel type, transmission type, number of seats and steering type.

**Machine learning Methodology:** The approach proposed is based on renew Machine Learning algorithms such as random forest regression, to predict the selling price of a used car based on key features.

Pseudo code:

Graphical user interface, text, application, email

Description automatically generated

**Experimental data**: The dataset was collected through Kaggle since it is an open platform to download any kind of datasets. This set includes more than 4250 records.

**Evaluation methods** : The regression models can be evaluated using the metrics such as Mean Sqaured Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These are the metrics that describes the accuracy of the performance between Existing and Proposed Systems.

**Results:** The data techniques were used such as encoding categorical variables and performing feature engineering activities. Among many Machine Learning algorithms, the Random Forest model, adjusts the dataset well and shows the accuracy about 90%.

1. **Data** :

Source of the data has been taken from Kaggle: <https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho> having data size of 1.41 MB

 This dataset contains information about used cars. This data can be used for a lot of purposes such as price prediction to exemplify the use of linear regression in Machine Learning. The columns in the given dataset are as follows:

1.     Name: Name of the cars

2.     Year: Year of the car when it was bought

3.     selling\_price: Price at which the car is being sold

4.     km\_driven: Number of Kilometres the car is driven

5.     fuel : Fuel type of car (petrol / diesel / CNG / LPG / electric)

6.     seller\_type: Tells if a Seller is Individual or a Dealer

7.     transmission: Gear transmission of the car (Automatic/Manual)

8.     Owner: Number of previous owners of the car.

1. **Data Exploration:**

Data Exploration is the initial step in the process of data analysis. During our work, we used correlation which explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

We used Pairplot to visualize the data to find the relationship between them where the variables can be continuous or categorical. Pairplot is a module of seaborn library which provides a high-level interface for drawing attractive and informative statistical graphics.

Heatmap is used to represents the coefficients to visualize the strength of correlation among variables. It helps find features that are best for Machine Learning model building. The heat map transforms the correlation matrix into color coding has been done during our work.

Below are the results of data exploration done in our project.

Final dataset:

Table

Description automatically generated with medium confidence

Final dataset correlation result:

Table

Description automatically generated

Pairplot:

Graphical user interface, text, application, email

Description automatically generated

Chart

Description automatically generated

Calendar

Description automatically generated

HeatMap:

Chart, treemap chart

Description automatically generated

Chart

Description automatically generated

1. **Data Preprocess and preparation:**

Data Preprocessing

It is the first and most important step in the process of developing predictive models. Pre-processing of data is a process of arranging the raw data that suits for machine learning algorithms.

a) Handling Missing Values

If there are any null values or missing values they should be ignored or replace null values with mean, median and mode strategies to handle the missing and null values.

b) Encoding Categorical Data

Since the prediction is done using Random Forest Regression, the non-uniform data fields are converted to integers ranging from 0 to 1 for accepting the parameters.

Dividing the data into test set and training set.

The dataset is dissected into two phases one is for training which is of 80% and the other is for testing which is of 20%. This is the convention that in any Machine Learning Algorithm the dataset is divided into two parts. Training phase describes the creating the model and the testing phase determines the bringing accuracy from the created model.

1. **References:**
2. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9696839>
3. <https://towardsdatascience.com/predicting-car-price-using-machine-learning-8d2df3898f16>
4. <https://arxiv.org/pdf/1803.11227v2.pdf>
5. <https://medium.com/odscjournal/predicting-car-prices-using-machine-learning-and-data-science-52ed44abab1b>
6. <https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho>
7. <https://quantdare.com/what-is-the-difference-between-extra-trees-and-random-forest/>
8. <https://pythonbasics.org/seaborn-pairplot/#:~:text=A%20pairplot%20plot%20a%20pairwise,axis%20across%20a%20single%20column>.
9. <https://seaborn.pydata.org/generated/seaborn.heatmap.html>
10. <https://ieeexplore.ieee.org/document/9835872>