**ABSTRACT**

In this paper we are investigating the application of supervised techniques of machine learning to “**Predict the Price of Used Cars”**. The forecasts are based on collected historical data from “Kaggle”and real time data from “Automax” in Windsor. Different methods have been used to make the predictions, such as Linear Regression Model,K-Neighbours Regression,Decission Tree Regression ,Random Forest, Gradient Boosted Regression Model, Ridge Regression, Lasso Regression ,Keras.Then, the predictions are evaluated.

**INTRODUCTION**

The need of transportation is a necessary demand. Everyone in the world rely on transportation to move from one point to another. A brand new car may not be affordable for everyone, so the alternative that everyone seeks is used cars. Cars are not a long-term investment, therefore the market of used cars has drastic impact in commercial market. As compared to brand new cars, the price of used cars are much more flexible .Manual prediction of used cars depend on the interest of vendors .In this scenario, the commercial value of price prediction of used cars comes into picture. As the car’s value depend upon various factors like year of manufacture, number of seats, engine power,brand value ,type of fuel,braking system,number of previous owners, odometer range, vehicle type, transmission type etc..the prediction is not an easy factor. Based upon the minimal details available it is important find the feature importance and also predict better accuracy .All the features listed above were not taken in consideration as some of them were irrelevant to the target variable. Five C’s on ethical use of data was taken into consideration which includes consent, clarity, consistency, control and consequences.

The objective of this project is to predict the price of used cars using machine learning models. Since it is a regression problem various regression models were analyzed like Logistic regression,decision tree, k-nearest neighbour, naïve bayes, Support Vector Machine and linear regression. The dataset consists of 6020 rows and 12 columns. Analysis were done on a real time dataset from “AUTOMAX”, which is one among the retail car dealers. The performance of all the various models were analyzed to choose the best out of them which would give a better prediction. The project focused on deep learning techniques using Kera’s library.

**RELATED WORKS**

1.Price Prediction Model for 2nd Hand Cars. (2020). Retrieved 24 February 2020, from <https://medium.com/@elysekatrina.go/price-prediction-model-for-2nd-hand-cars-f1801d8c8d47>

**Elyse Go(2019)**: Here the main heading of their project is “ We Made a Price Prediction Model for Used Cars in the Philippines “. In this project they were trying to predict the price of used cars by creating an online prototype application just like an online calculator. They train their model with the data that they have collected in Philippines. So in this article they haven’t mentioned much about their dataset, only thing which they said was that they were using 12 features which was different from other 2 –3 features from other online calculators. In their dataset they were having null values in the column “color family”, “fuel type” and “mileage”, in the “color family” and “fueltype” column they filled the null values with the value “ No color ” and “ No fuel type “ and they filled the null values in the “mileage” column by imputing the values using K-Nearest Neighbors Regressor. So in model building they used cross validation technique with 5 folds n they used Decision tree model (73.45%) , XGBoost (79.63%) and Random Forest (80.14%). So they got the highest accuracy for Random forest model.

2. How much is my car worth? A methodology for predicting used cars prices using Random Forest (2020). Retrieved 24 February 2020, from <https://arxiv.org/ftp/arxiv/papers/1711/1711.06970.pdf>

**Future of Information and Communications Conference(2018)**:Random Forest Method was used to predict the price of used cars based upon different features. Initially the data was collected, and feature importance was identified. The correlation between various features to price was analysed. Then data was pre-processed and the model was applied with features as input and price as output. The data was about 370000 used cars. In exploratory analysis ,visualizations were done to understand the difference in data. Since it is a regression problem, linear regression model was also tested. Then it was found Random Forest solved over fitting problem. Regression accuracy was less than 75% in training data. The dataset resulted in an accuracy of 83% for test data and 95% for train data.

3. Vehicle Price Prediction System using Machine Learning Techniques. (2020). Retrieved 24 February 2020, from

<https://pdfs.semanticscholar.org/8823/297bb2d789f98151bad70317ea97c1fc6fac.pdf>

**Kanwal Noor(2017)** -Concluded in this research developing a good regression model offers accurate prediction of car price in order to do this they have considered price as the independent variable and other attributes as dependent variable and the data was processed using multiple linear regression technique for price prediction. Since the engine type and model was textual they have used Minitab to convert into category codes(1,0) because regression works on numeric data. As said, will try to implement advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices.

4. Predicting the Price of Used Cars using Machine Learning Techniques (2020). Retrieved 24 February 2020, from <https://s3.amazonaws.com/academia.edu.documents/54261672/2014_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Techniques.pdf?response-content-disposition=inline%3B%20filename%3DPredicting_the_Price_of_Used_Cars_using.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20200224%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20200224T043332Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=bc82d5703e5a4d3d4a734bda96f573dc3a27ce5b77c707862a786e41f6805afb>

**Sameer Chand Pudaruth(2014**) - In this paper they have used supervised learning models to predict the price of the used cars in Mauritius .The data was collected from the daily newspaper and different techniques like multiple regression analysis -nearest neighbour, naive Bayes ,and decision tree was used to make prediction .Then the algorithms are evaluate and compare to find the best performance .Therefore in our project we have collected the real time data and will try to implement advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices.

5. Predicting the Price of Second-hand Cars using Artificial Neural Networks (2020). Retrieved 24 February 2020, from <https://www.researchgate.net/publication/319307014_Predicting_the_Price_of_Second-hand_Cars_using_Artificial_Neural_Networks>

**Saamiyah Peerun, Nushrah Henna Chummun and Sameerchand Pudaruth (2015):** The objective was to find the price of used cars using Artificial Neural Networks, So the data from 200 was analysed worked on four different machine learning algorithm. It was found Support Vector Machine Learning produced comparatively better results than Neural network and Linear Regression, But for highly priced cars the predicted values was higher than actual prices. So further analysis is required using more data to get better predictions. In Section 1 ,it is the comparison of the demand of used cars from 2003 to 2014.In Section 11, different results on neural networks is summed up. In Section 111, data collection and methodology is explained. Section 1V describes the result of Price Prediction. In conclusion part, there is suggestion of ideas for future works. Four different models which are KNN, SVM, Linear Regression and Decision Tree was analysed. It was found Neural Network with 1 hidden layer and 2 nodes had smallest MAE. Support Vector Regression had predictions which was better than Linear Regression. KNN had gave less accuracy model. Everything was analysed using cross- validation value of 10 folds.

6. Predicting Used Car Prices with Machine Learning Techniques. (2020). Retrieved 5 March 2020, from <https://towardsdatascience.com/predicting-used-car-prices-with-machine-learningtechniques-8a9d8313952>

**Raschka, S., & Mirjalili, V. (2017):** By performing different models, different viewpoints were explored, and their performance was eventually compared. The aim of this analysis was to predict prices of used cars using a dataset that had 13 predictors and 380962 observations. The dataset was discovered with the support of data visualizations and exploratory data analysis, and features were explored in detail. This explored the relationship between features. Predictive models were implemented at the last stage to forecast car prices in an order: random forest, linear regression, ridge regression, lasso, KNN, XGBoost.

Taking all four metrics into consideration, it can be concluded that random forest is the best model for forecasting used car prices. Random Forest provided the best MAE as a regression model .