Price Prediction of Used Cars using Deep Learning

Group 7:

Amith John Varkey Abdul Hannan

Delta Joseph

Abstract

The need of transportation is a necessary demand. Records from previous years depicts that number of people relying on used cars has been increasing. This paper deals with prediction of used cars using deep learning techniques. Records of sales of used cars from 2012 was **doubled** to its initial sales in just six years in 2018.Traditionally car dealers used manual prediction in order to predict the price of used cars to a customer. The main drawback to this approach was that price would differ from one car dealer to other car dealers since it's based on random assumption. So we thought of implementing this project which would predict a standardized price of the used cars. Now the application of Machine Learning comes into picture where it is capable of handling a large amount of structured and unstructured data. Therefore by this approach the basic price of pre-owned cars would be universal.

1. Introduction

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 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 eight regression models were analysed like Linear regression, K-neighbour regression, decision tree, random forest model, gradient boosting, ridge regression, lasso regression ,XGboost regressor .Artificial Neural network using keras package was also analysed on the data. The main dataset which we got from 'Kaggle'

consisted of 6020 rows and 12 columns. Analysis was also done on a real time dataset from "AUTOMAX", which is one among the retail car dealers in Windsor, ON. The "Automaxx" dataset contained 200 rows. The performance of all the various models were analysed to choose the best out of them which would give a better prediction. The project focused on deep learning techniques using Keras library. The data was explored by splitting the dataset into test ratio of 0.2 and 0.25 to analyse the difference.

2. Related Works

[1] **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 was declared that they were 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 "fuel type" 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 using K-Nearest Neighbors Regressor. They used label encoder to transform the categorical values into numerical. So, in model building they used cross validation technique with 5 folds and they used Decision tree model (73.45%), XGBoost (79.63%) and Random Forest (80.14%).In conclusion ,the highest accuracy and the best model was Random forest model. To see their idea come to life, they had to create a web application which could be hosting and integrated into the client's website.

[2] **Future of Information and Communications Conference(2018)**: The title of the project was "How much is my car worth? A methodology for predicting used cars prices using Random Forest." The data was about 370000 used cars. Initially ,identified the important features that reflect the price. Then they have pre-processed the data to removed null values and least important feature that does not help in model building.

Random Forest Method was used to predict the price of used cars based upon different features. and the model was applied with features as input and price as output. In exploratory analysis ,visualizations like bar chart box plot, distribution graph etc. were used 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. This overcomes by averaging out the predictions of individual trees with a goal to reduce the variance and ensure consistency. 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. The most relevant features used for this prediction are price, kilometre, brand, and vehicle Type by filtering out outliers and irrelevant features of the dataset.

[3] **Saamiyah Peerun, Nushrah Henna Chummun and Sameerchand Pudaruth (2015):** The objective was to find the price of used cars with the project title "Predicting the Price of Second-hand Cars using Artificial Neural Networks". So the data from 200 cars from different sources was gathered and analysed on four different machine learning algorithm. 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. 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 conclusion, the purpose of this paper was to forecast the price of recycled and second-hand used cars in Mauritius and suggestion of ideas for future works were declared. Finally, 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.

[4].**Raschka, S., & Mirjalili, V. (2017):** In this paper the project was named Predicting Used Car Prices with Machine Learning techniques. This project was consider to solve the predict the price of the car with traditional method rather than going for online services which may not be the best. As it is important to know the actual price of the car while both selling and buying. The data used in this project was downloaded from Kaggle called craigslist used cars data file. In data cleaning un-necessary features were removed like url','image_url', 'lat', 'long', 'city_url', 'desc', 'city', 'VIN' features were dropped totally. 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 .

[5].**Stefan Lessmann**: Concluded in the research paper called Resale Price Prediction in the Used Car Market, that a previous work was done with very few studies which had explicitly attempted to predict resale prices with maximal accuracy to support decision making. As a consequence, answers to the following questions were unclear : i) to which degree are resale prices predictable, ii) what is the relative accuracy of different prediction methods and are some methods particularly effective, iii) given that market research agencies have specialized in residual value estimation, is it sensible for car makers to invest into an in-house resale price forecasting system? The objective of this paper is to provide empirical answers to these questions. using real-world sales data from a leading German car manufacturer, a largescale empirical benchmarking study is undertaken to contrast the predictive power of alternative prediction methods. And the private information was only available to used cars makers/vendors.

They concluded by saying , empirical results suggest that the Ensemble Selection (ES) methodology performs best in resale price forecasting. Using this approach, the mean absolute error (MAE) of forecasts is only 3.97 but random forest performs(RF) as good as ES. Finally they suggest that the methods most widely used in resale price modelling are least effective. Linear regression methods predict significantly less accurately than advanced methods such as RF and ES.

[6] **Sameer Chand Pudaruth(2014**) - In this paper the project was named as "Predicting the Price of Used Cars using Machine Learning Techniques" in Mauritius. The predictions are based on historical data collected from daily newspapers. Initially, it collected 400 + data. For example, after further pruning, they kept only the three most common make in Mauritius, i.e. Toyota, Honda, Renault, and etc... They eliminated all the makes for which less than 10 records existed. About the volume of the cylinder, this was offered in a range for certain vehicles.

Different techniques like multiple regression analysis K-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 the project they 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.

3. Methodology

In this project, we have used 8 different supervised regression models like Linear Regression, K- Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regression, Ridge, Lasso, XGBoost Regressor and also Artificial Neural Network using the Keras package in order to predict the car prices.

Linear Regression : A linear regression is used in order to find out if there is any linear relationship between the dependent variable and the rest of the independent variable.

K- Neighbors Regressor : It is an algorithm which takes in all available values and predict the output based on some similarity benchmark (e.g., distance function).

Decision Tree Regressor : This algorithm consist of a tree like model and will make predictions by asking a set of questions or decisions to the model.

Random Forest Regressor : This algorithm consist of multiple decision trees and it will join the predictions together to give more accurate prediction rather than depending on an individual decision tree.

Gradient Boosting Regressor : This algorithm also tries to improve the accuracy of the decision tress, Instead of training a single tree multiple trees are trained in the sequel in order to improve the result.

Ridge Regression : This is an algorithm which uses L2 regularization and is also used for data that suffers from multicollinearity. Here a parameter called 'alpha' is used as a smoothness constraint.

Lasso Regression : This is an algorithm that uses L1 regularization and it performs both variable selection and regularization in order to improve the prediction accuracy of the model.

XGBoost Regressor : XGBoost regressor is a decision tree based machine learning model which uses a gradient boosting structure.

Artificial Neural Network (Keras) : Artificial neural network performs similarly to that of a human neuron. So this algorithm learns itself from the input and then gave us a result. It consist of 3 layers: input layer, hidden layer and the output layer. Keras is an open-source neural network library that is used in python.

So before getting into modelling we have done data exploration to understand more about the data and have also done data cleaning in order to make the data ready for modelling. In data exploration we have analysed for different features in our data like 'Fuel Types', 'Transmission Types', 'Owner Types' and also 'Car Company'.

Figure 1

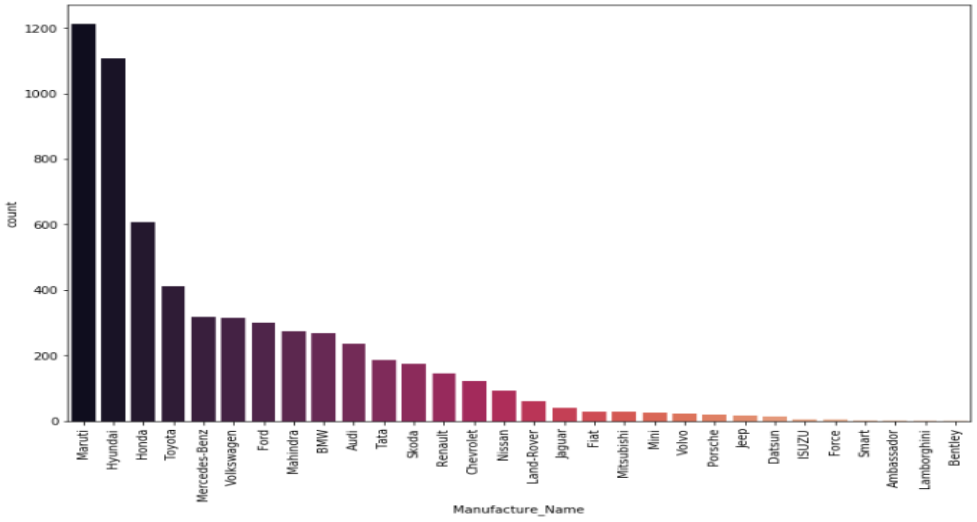

Figure 2

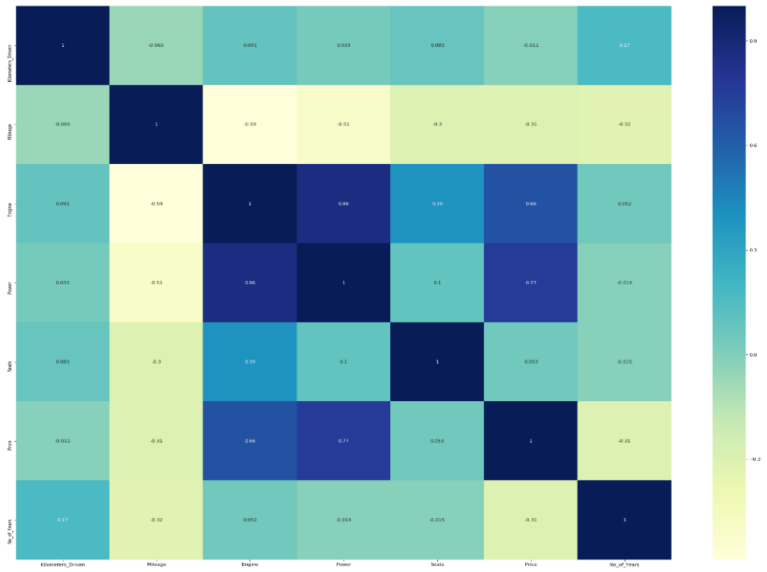

Figure 1. is a histogram showing how many records are there for the different Car Companies that we have in our data and Figure 2 is a correlation plot showing how different features in our data are correlated with each other. In this we can clearly see that the feature 'Kilometers_Driven' is having less correlation value of 0.011 with the price column (which is the target column in the project). So by looking at the value we can say that we will drop the 'Kilometers_Driven' column but since we know kilometres driven is an important attribute which is required to predict the price we will not drop that column for modelling.

In data cleaning we have imputed mean and mode values for the null values that we have encountered in the features. We have also converted the categorical values into dummy values by using the get_dummies function in python, since we know that machine learning will only work for numerical values and not for textual data. We also created two more columns one which shows the number of years the car was manufactured and other column which had the car name. We also dropped column which were of least importance. In data modelling part we have split the data with a test size ratio of 0.2 and 0.25 and have performed modelling with these two test ratios. We also scaled the data after splitting by using the MinMax Scaler function to make the values in the range of 0 to 1, since we were having different range of values for the feature.

We then carried out the same set of procedures that we have done before and after modelling in the Kaggle dataset for the real time dataset that is the Automaxx data.

4. Results

So after carrying out all the data exploration, pre-processing and by applying the various regression models and also ANN, we evaluated the model by using the two metrics **MAPE (Mean Absolute Percentage Error)** and also **RMSE (Root Mean Square Error)**. We can't say that these are the best two metrics for evaluation , but in this project we have used these for evaluation of the model. Table 1 shows the values for MAPE, RMSE and also the R2 score for test size =0.2.

Table 1

Model	MAPE	R2_Score	Rmse value
Linear Regression	46026261912.8	-5037658137424.00	141785128332229.9
K-Neighbor Regression	18.7	0.87	7295.26
Decision Tree Regression	19.6	0.86	7536.69
Random Forest Model	15.8	0.92	5778.68
Gradient Boosted Regression	54.4	0.38	15675.95
Ridge Regression	26.1	0.83	8250.51
Lasso Ridge Regression	25.9	0.82	8441.84
Artificial Neural Network(Keras)	15.3	0.91	5974.51
XGBoost Regressor(k=5)	15.8	0.89	6580.26

Table 2 shows the values for MAPE, RMSE and also R2 score for test size =0.25. Here XGBoost regressor with K value as 5 was having the best values.

Table 2

Model	MAPE	R2_Score	Rmse value
Linear Regression	47701943920.3	-18146552	274771265431764.22
K-Neighbor Regression	19.2	0.84	8034.26
Decision Tree Regression	20.1	0.81	8953.43
Random Forest Model	16.1	0.89	6789.01
Gradient Boosted Regression	51.0	0.43	15460.90
Ridge Regression	25.8	0.83	8492.17
Lasso Ridge Regression	25.6	0.81	8956.91
Artificial Neural Network(Keras)	15.6	0.89	6903.90
XGBoost Regressor(k=5)	15.2	0.90	6424.22

Table 3 shows the values for MAPE,RMSE and also R2 score for modelling that we have done for the real time data that is the Automaxx data. Here we have done for test size = 0.2.

Table 3

Model	MAPE	R2_Score	Rmse value
Linear Regression	67822346	-79005870927	4695386783326496.0
K-Neighbor Regression	16.3	0.54	3595.89
Decision Tree Regression	16.0	0.54	3567.74
Random Forest Model	12.2	0.73	2745.98
Gradient Boosted Regression	18.2	0.44	3935.66
Artificial Neural Network(Keras)	12.1	0.74	2705.47
XGBoost Regressor(k=5)	11.8	0.76	2612.21

5. Discussion

The main challenge that we faced was to collect real time data because of confidentiality. Most of the car dealers we visited said that they are not ready to share their data because of confidentiality. But we got a positive response from " Automaxx Windsor" which was one of the car dealers in Windsor, ON. In this data we were not having enough amount of rows to

work on, so we used one of the dataset from “Kaggle” which we used as a main data for analysis and also did modelling on the real time data.

We have done modelling on the main data using different regression models mentioned in section 3 and also by using Artificial Neural Network for 2 different test sizes in order to find out which model is having better result and also for what test size. Here we have used test size of 0.2 and 0.25. So for test size = 0.2 we were getting better result for **ANN (Artificial Neural Network)** since it was having MAPE value as 15.3 and RMSE value as 5974.51 which was the least value among the rest of the models that we have tested for test size of 0.2 and linear regression was giving us the worst result. R2 score for ANN was 0.91, whereas for Random Forest was 0.92. But the MAPE and RMSE values were favourable for ANN. For test size = 0.25 we were getting better results for **XGBoost Regressor** with number of folds i.e. k =5. The MAPE value was 15.2 and RMSE value was 6424.22 for XGBoost regressor which was the least among the rest of the models that we have used for modelling and here also linear regression was giving us the worst result. Even the R2 score for XGBoost regressor was 0.90 which was better among the rest of the models for test size =0.25.

We also performed modelling using the different regression models mentioned in session 3 and also Artificial Neural Network for the real time data with test size = 0.2**. XGBoost regressor** with k=5 was giving us better values for MAPE and RMSE which were 11.8 and 2612.21 respectively whereas linear regression was giving us the worst result. The R2 score for XGBoost regressor was also higher than the rest of the models which was 0.76.

6. Conclusion

ANN was the best among all the models with test size =0.2 whereas XGBoost regressor was the best one with test size = 0.25. But when we compare both the results from test size =0.2 and 0.25 for our main data, **Artificial Neural Network (ANN)** was performing better than XGBoost regressor with k=5. The reason why we have came to this conclusion is that, even though the MAPE value for ANN was 0.1 higher than the value for XGBoost regressor, the RMSE value for ANN was 5974.51 whereas for XGBoost regressor was 6424.22 which was 449.71 higher than the value for ANN. Also the R2 score for ANN was 0.01 more than XGBoost regressor. So we can come to a conclusion that ANN was the best model which will help us predict the car prices based on historical data that we have worked on. In most of the research papers Random Forest was the best model, but for us ANN was performing better. We can improve the accuracy of the model in the future by using fuzzy logic and also genetic algorithm to predict the car prices.

For the real time data, **XGBoost regressor** was performing better than ANN since XGBoost regressor was having the best MAPE, RMSE and R2 score values. In all the 3 parts that is for modelling with test size =0.2, with test size =0.25 and also for modelling with the real time data, Linear regression was giving us the worst values maybe because there wasn’t any linear relationship between the dependent variable(i.e. ‘ Price ‘) and rest of the independent variables.

7. Contributions

Amith worked on data cleaning of the main data. He worked on different models like Random Forest Model, Ridge, Lasso and also Artificial Neural Network for all the 3 parts that is for modelling with test size =0.2, modelling with test size =0.25 and also for modelling with the real time dataset. He also worked on report writing in "Methods" , "Discussion" and "Result" session.

Abdul worked on data exploration part of the main data. He worked on different models like Gradient Boosting regression, Decision Tree and also helped Amith with Artificial Neural Network for all the 3 parts. He also worked on report writing in "Related Works" and "Conclusion" session.

Delta worked on data exploration and data cleaning of the real time data. She worked on different models like Linear Regression, K- Neighbor Regression and also XGBoost regressor for all the 3 parts. She also worked on report writing in "Abstract" and "Introduction" session.

8. References

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.
2. Pal, N., Arora, P., Sundararaman, D., Kohli, P., & Palakurthy, S. (2020). How much is my car worth? A methodology for predicting used cars prices using Random Forest. Retrieved 8 April 2020, https://www.semanticscholar.org/paper/How-much-is-my-car-worth-A-methodology-for-used-Pal-Arora/6e2377d6ca202d9ce8a7bcb229cd7026309962e4
3. 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
4. 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
5. Resale Price Prediction in the Used Car Market (2020). Retrieved 8 April 2020, from https://pdfs.semanticscholar.org/edd2/191dae1b2f10052c51fd61a908e0b91fbb74.pdf
6. Predicting the Price of Used Cars using Machine Learning Techniques.(2020). Retrieved 8 April 2020, from https://www.ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf
7. scikit-learn: machine learning in Python — scikit-learn 0.22.2 documentation. (2020). Retrieved 17 April 2020, from https://scikit-learn.org/stable/
8. Home - Keras Documentation. (2020). Retrieved 17 April 2020, from https://keras.io/
9. Expanding your machine learning toolkit: Randomized search, computational budgets, and new algorithms. (2020). Retrieved 17 April 2020, from https://cambridgecoding.wordpress.com/2016/05/16/expanding-your-machine-learning-toolkit-randomized-search-computational-budgets-and-new-algorithms-2/
10. Digitaltrends. (2020). Retrieved 17 April 2020, from https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/