USED CARS RESIDUAL PRICE ESTIMATION

Batlle Lalande, Jordi
School of Computing
National College of Ireland
Dublin, Ireland
x17133246@student.ncirl.ie

***Abstract* – Over the last years, the used cars market has shown an important growth. This paper proposes a regression model to allow both the car industry and the private sellers to estimate the best reselling price of used cars. This paper assesses as well the determinants that contribute to determine that price.**

The research was built from 3 datasets of used cars from Australia, Germany and Spain scrapped from the web. After some preprocessing tasks, these datasets were used to compare the performance of 5 different regression models. Among the models used, it was found that for this data linear models provided a good performance and little difference was found between Multiple, Ridge and Lasso regressions. Then, decision tree was used and showed a very poor performance. Finally, the best performance was found when Random Forest was used. This model, compared to the second-best model, reduced the RMSE by 15% to 20%.

Keywords — Used cars pricing, Forecasting, Regression models comparison, Linear Regression, Decision tree, Random forest

I. INTRODUCTION

Over the last years, after the financial crisis, there has been numerous changes in the economy worldwide. One of the sectors that shows a deep change is car sales. The difficulty to buy new cars has evolved in an increasing demand for used cars, which sales in the United States has been increasing year to year since 2013[1](fig.1). At the same time, the rise of new business models like Uber or Cabify has contributed to increase the stock of used cars. In parallel, car leasing business, another key factor for the used cars stock, has seen a high rate growth worldwide[2] (fig.2) and it is estimated the market size will grow 18.38 million units during the period 2019-2024, a 15% compound annual growth rate (CAGR)[3].

This used car growing industry has driven a necessity to build models to understand the market; in particular to evaluate the features customers will consider most relevant when they look at buying a used car, which may be different from new cars, and to set a fair price both from the seller and the buyer perspective.

Price car estimation models are not new but, generally, these old models rely on a linear relation between price and the car features [1]. At the same time, over the last years, there has been important changes in car models, particularly with the introduction of hybrid and electric cars. And of course, the machine learning field has evolved notably. Due to these changes, recent years have seen a rise of research in the domain and several authors have published their model proposals in order to improve traditional models using more modern data mining techniques.

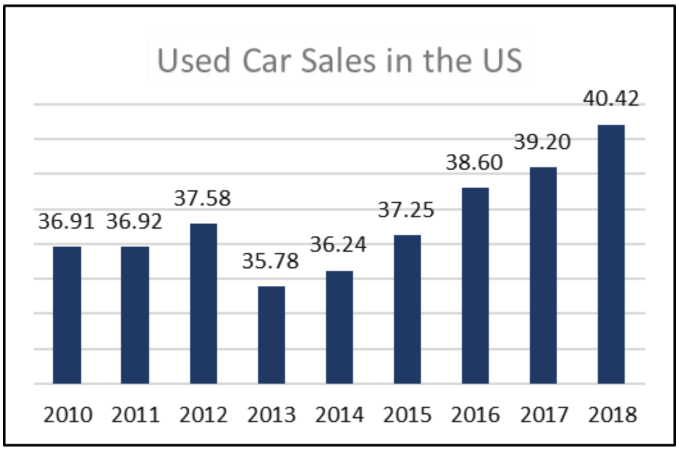

Fig. 1. US used light vehicle sales 2000-2018 (in millions)

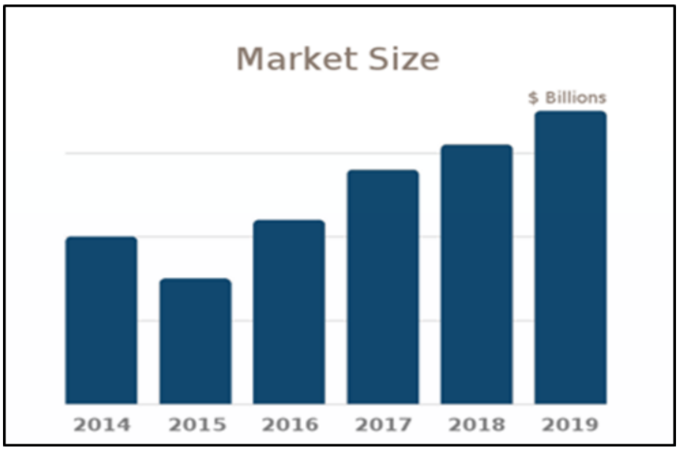

Fig. 2. US market size for used cars ($billions)

[1] www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990
[2] www.anythingresearch.com/industry/Used-Car-Dealers.htm
[3] www.technavio.com/report/car-leasing-market-industry-analysis

The objective of this research is to find out what are the most relevant attributes in order to set used cars price and to build a model capable to predict the fairest price of any used cars based on its characteristics.

It is expected that the resulting model will benefit all the agents involved in the market:

- Car dealers: the model should allow businesses to understand consumer preferences; then they will be able to set prices to maximise business profitability.
- Private sellers: for individuals who are selling their car, it will be important to estimate accurately the price in order to close the deal quickly while receiving a fair price for their cars.
- Leasing companies: the model is relevant for them in order to estimate accurately the cars residual price at lease term; under evaluate the residual price will turn in a loss while over estimation will reduce the number of clients who will perform the buy and, as a result, the company will incur in charges such as restocking or advertising to sell the car in the used car market.
- Used car estimation websites: the model will help to improve their tool accuracy. These websites are usually free and base their profit on online advertising, therefore better estimations should attract more users and therefore increasing advertising profits.
- Buyers: the final beneficiary of this model are of course the individuals looking to buy a car as they will be able to compare offers and take more informed decisions to ensure they are paying the right price.
- Banks/lending agencies: better price estimations should incur in better borrowing assessment.

This paper is organised as follows. In the next section, a review of related work is provided. Section III provides details of the datasets used in the research and the methodology applied to process the data; then section IV goes through the different data mining techniques used to predict used cars price and provides results evaluation of these techniques. The last section includes a summary of the research and its results and a critical assessment of the work done.

II. Related Work

As mentioned previously, used car price estimation is not a new problem, but research has experienced a recent increase due to the necessity to upgrade models to new market circumstances.

During the years 2000's, some researchers looked at the determinants of used cars price. Cho, in an article in the Journal of Economic Research [2], found out that car age and mileage contribute negatively in the resale price, that brand image has a positive impact and that prices were influenced by seasonality, with prices starting to increase in Spring and August being the peak season. In his research, Cho used only rental cars which, due to the nature of the business, are very well maintained and, therefore, it might be erroneous to extrapolate the results to random used cars.

A similar approach was used by Richardson in his thesis [3]. Richardson based his work on the hypothesis that car manufacturers are prone to build cars which do not depreciate quickly; in particular, he used a multivariate linear regression model to show that hybrid cars do not depreciate as quickly. Whether the car is hybrid or not would then be an important determinant for used car price prediction.

Another research by Andrews and Benzing studied the factors that influence the price in eBay's used car auctions [4]. They used a binary logit model to discover that seller reputation and clarity of the title had greater selling chances. This research is very limited from the methodology point of view, the research market used is narrow and very specific as opposite to the objective of the present work.

Other researchers looked not only to variables but also to build a price estimation model. Kuiper, in an article published in the Journal of Statistics Education [5], used a multivariate regression model to determinate used cars price based on its characteristics. Although the limit of her work comes from the fact that only 800 used cars records were used, all of them from General Motors with less than 1 year and in excellent condition.

Du et al. released a paper based on their experience working at Power Information Network. The objective was to provide a solution to leased cars returned at the term of lease, as in most occasions, clients do not execute the car buying clause. They developed the Optimal Distribution of Auction Vehicles (ODAV) system [6] to estimate the best reselling price and when and where to resell the cars; ODAV is an automated system to help sellers to maximise profits. The system integrates 3 models: kNN regression for price forecasting, autoregressive moving average time series to determine the volume elasticity against the price and optimisation algorithm for car distribution.

Wu et al. used an adaptive neuro-fuzzy inference system (ANFIS) approach to used cars price forecasting [7]. They found it performed better than conventional artificial neural network (ANN) with back-propagation. Even though their model performed well, it is based uniquely on 3 characteristics: make of the car, manufacturing year and engine style.

Listiani, on her Master thesis [8], found that, when large data sets are available, SVM regression provided better results than linear and multivariate regression to predict leased cars residual value. This technique proved a better handling of high dimensional data and avoids the under-fitting and over-fitting curse. On her approach, Listiani opted for not removing outliers; decision that can distort the regression model.

Gongqi et al. [9] proposed a model based on a combination of BP neural network and nonlinear curve fit to predict used cars residual value. The advantage of this model is it provides a good fit for non-linear relations. Their research concluded that mileage, manufacturer and estimated useful life were the main attributes to fix the price.

In more recent years, new, more advanced approaches were used in the domain. Pudaruth used multiple linear regression, kNN, naïve Bayes and decision tree techniques to predict used car prices [10]. Unfortunately, the researcher did not find high accuracy rates on his model certainly because after pre-processing the data, only 97 records were retained and variables with high rate missing data deleted.

One year later, Pudurath participated in another research [11] in which the researchers compared the performance of 4 regression methods: Support Vector Regression (SVM), Linear Regression, k-Nearest Neighbour (kNN) and multi-layer perceptron with back-propagation. Out of these models, kNN provided the worst results with the other 3 providing similar Mean Squared Errors (MSE). As with the previous work, the sample in this case was only of 200 records and out of the initial variables, only 6 were retained (manufacturing year, make, model, mileage, horsepower and country of origin).

Lessman and Voß [12] found that random forest regression performed better than linear regression models with high dimension models. They found as well presence of heterogeneity in price forecasting and that private dealers have a market knowledge advantage over market research agencies that allow them better price predictions and therefore agencies should consider investing on in-house estimators rather than rely on external sources.

Noor and Sadaqat used a multiple linear regression model [13]. Their model was built from only 1699 records and although initially they had 16 variables, it was reduced to just 3 of them: year, model and engine type. Based on most of the works published, it is considered that a variable such as odometer should play an important role in used cars price prediction and that other techniques which do not assume a linear relation between price and the independent variables should be considered.

Zhang et al. used classification algorithms to model price prediction of used cars [14]. Out of the methods used (Logistic regression, SVM, Decision tree, Extra trees, AdaBoost, Random forest), they found out that random forest showed the best performance. Although, it must be said they only used 5 variables to build their models. Also, their approach only used classification methods which leads to lower precision in price prediction as it generates predictions on price ranges.

Pal et al. built their price estimation model on the grounds that used car sellers are taking advantage on the increasing demand to inflate used car price [15]. They used random forest for their model which turn in high accuracy rate. Although, to build their model they only used dealers' prices and dropped private sellers records. Also, the drop the location variable that was present in their initial dataset. As seen in other models, location variable can be important in building the model.

Go made a price prediction model for used cars in the Philippines [1]. One particular aspect of her model is that it included the car retail price as new. As a result, the model which included 12 independent variables showed that only 2 variables explained 87% of it, and a relevant variable for every other model as mileage only accounted for 3%. The issue in this case is that in real-world market getting to know the initial price is very difficult and it depends on a few factors, as equipment for example, and that seller may be reluctant to provide this information.

Finally, in recent research published early 2020, Gokce compared 6 methods (Random Forest regression, Linear regression, Ridge Regression, Lasso, K-Nearest Neighbor and XGBoost) to estimate used cars price [16]. He found that random forest regression model provided the best results and that year, mileage, make, drivetrain type, fuel used, manufacturer and number of cylinders were the most important attributes to determinate the price. Gokce, during the pre-processing, dropped extreme values for price and mileage both from the lower and upper end but in terms of manufacture year, he only dropped values for cars earlier than 1985. The present research considers that further work needs to be done on the year of manufacturing as the gap from 1985 to 2019 used by Gokce is too large. In particular, the first decision node of his decision tree separated records between less than 0.8 years old and more than 0.8 years old.

III. METHODOLOGY

Knowledge Discovery in Databases (KDD) methodology was used to perform the research, which is a 5-step process widely used in machine learning. Once the objective of the research was set, the first step of the process consisted in selecting the data. For this, 3 datasets were downloaded from Kaggle repository. They contained information of used cars on sale in Australia[4], Spain[5] and Germany[6]; the data included variables such as price, year, mileage, brand, model, fuel type, transmission type, engine, type of the car among others.

The second step was the data pre-processing, probably the most challenging part of the process. A few steps were carried out once the data was loaded. The process started by removing irrelevant variables for the research as for example id, URL, ad ids and description. Then, set the data types; in particular, categorical variables were coded as factor type variable. Afterwards, some data exploration was carried out and variables looked at one by one. At this stage, boxplots and histograms were used to look at the numerical variables distribution and detect deviated values. From these values, minimum and maximum value thresholds were stablished and

[4] https://www.kaggle.com/bahamutedean/secondhand-car-price-estimation
[5] https://www.kaggle.com/harturo123/online-adds-of-used-cars
[6] https://www.kaggle.com/bozungu/used-cars-listing-from-ebay

rows outside the range dropped. In particular, for the target variable, price, it was checked that no missing values were present then a lower limit of 500 was stablished for all datasets and the upper limit to equal the sum of mean and the standard deviation. Similar approach was used for the 2 most important predictors: registration year and mileage. Special attention was put to ensure that all values made sense. Then, for each categorical variable, plots were used to check the count of each factor in the datasets and factors that were outside the scope of the research eliminated; for example, new cars and buses were eliminated. Next, missing values were looked at. First, rows with more than 1 missing value were dropped, then k closest neighbours (kNN) algorithm was used to predict missing values within the numerical variables. Finally, all rows with remaining missing values eliminated (fig.3).

Spanish used cars, it was found that prices vary on the type of fuel they use, from cheapest to most expensive: gas, diesel, hybrid and electric cars.

Then, correlation between the main numerical variables was looked at (fig.4). It was found that price has a stronger correlation with the registration year than with the mileage. As well, for the Australian cars and the Spanish cars, strong relation was found between registration year and mileage while in the German cars it is much weaker.

Next step consisted in data transformation. During this step, one-hot encoding was used to transform each factor of the categorical variables into binary variables so they can be computed later by the machine learning algorithms.

Fig. 3 Datasets structure after the preprocessing stage

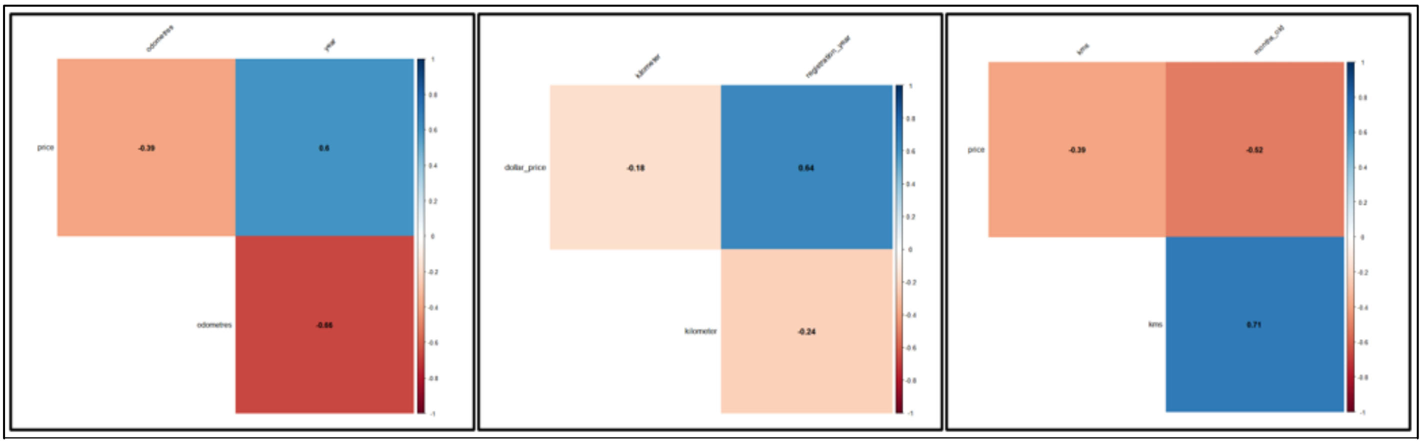

Fig. 4 From left to right, Australia, Germany and Spain correlation plots

At this stage, some plotting was done to have a visual representation of the price mean behaviour towards the different factors included in the datasets. In general, it was found that automatic cars are more expensive. On the German dataset, it was found that cars that advertised a discount were more expensive and that cars sold by dealers were as well more expensive than those sold by private users. For the

Numerical variables, except price as the models target variable, were transformed as well so they fit within the range between 0 and 1 as it is the case for the factors encoded previously.

The fourth stage of the KDD was to apply the selected data mining techniques for regression as the objective of the research is to predict the price of used cars. But previously, the data were split into train and test data frames and 5-fold cross validation set to train the data included in the train data frame. It is expected that the use of cross-validation in the train split will improve the final model. This technique was used in all data mining algorithms except Random Forest were the cross-validation was not applied to the train data frame but instead the model was built using 150 trees.

The final stage of the KDD relates to the evaluation of the results obtained through the application of the data mining techniques. In order to assess the performance of regression methods, Root Mean Square Error (RMSE) was used as it is a common measure to all techniques and it uses the same units than the estimated variable. RMSE basically consists in calculating the average of the squared difference between the model predictions and the actual values. Additionally, the coefficient of determination R-squared is used to compare the linear models performance. Although R-squared applies to linear models, from the combination of R-squared results and RMSE a model evaluation can be inferred. For example, if the RMSE result of Random Forest is better than the one of a linear model and that linear model has a high R-squared then it can be concluded that the Random Forest model is a good fit to the data. Otherwise, RMSE alone will only allow to compare the results between the different models but it does not provide information of the goodness fit.

Further details on the data mining process and the results obtained can be found in the next sections of this paper.

IV. EVALUATION

In this section, a short overview of the data mining techniques applied is provided along with the results found.

A. Multiple Linear Regression

The first algorithms this research applied are based on linear relation between price and the different predictors included in the 3 data frames.

Firstly, multiple linear regression was tested. As indicated previously, older researches relied on this technique to established price prediction models of used cars and therefore it can be used as a benchmark to compare the overall results. This technique relies on an existing linear relation between price and the year and mileage variables although it was found that the strength of the relation differed among the 3 datasets. Additionally, the remaining variables for each dataset were included in the model to test its performance and evaluate their importance.

When running the model in the test sample of the Australian used cars dataset, an RMSE of 3183 and an R-squared of 0.87 was found. Using the German used cars, RMSE of 1638 and R-squared of 0.73 were found. Finally, on the Spanish used cars, the results found were RMSE of 2434 and R-squared of 0.86.

In linear regression methods values close to 1 represent a good fit and this was the case. A high R-squared was found, even though the result for German cars is not as good, and therefore the RMSE can be used to compare the performance of other models.

Considering the factors, it was found that registration year and mileage were by difference the most determinant in all 3 models.

B. Ridge Regression

The second method used in this research is Ridge Regression. It is as well a multiple linear model, but it has the particularity it uses "shrinkage" method which consist in shrinking the coefficient of the predictors towards zero. It is commonly used to address issues when there is collinearity in the data. Although no strong collinearity was found, this method was used to contract its results with the other 2 linear methods used.

The following results were found. For Australian cars, the RMSE found was 3221 and Rsquared 0.87. For German cars, 1648 and 0.73 and for the Spanish ones, 2447 and 0.86.

In terms of factor importance, results differed with respect to the first model. Although year registration kept being on top, mileage was not anymore a relevant factor. It was found that for Australia, some types of engine were more relevant while for Spain and Germany it was the car model.

C. Lasso Regression

The next method used was Least Absolute Shrinkage Selector Operator (LASSO) regression, which is another method that uses shrinkage but, as opposite to the Ridge regression, it excludes the variables with coefficients equal to zero. As a result, a subset of predictors is obtained to improve the model accuracy. This method is particularly effective when the data has many variables. In this case, transforming the categorical variables into dummy variables increased the mode dimension and therefore it makes an interesting method to contrast the results with the other models.

The application of this model to the test data produced the following results. For the Australian cars, an RMSE of 3197 and R-squared of 0.88 were found. For the German cars, the RMSE was 1636 and the R-squared 0.73. Finally, the Spanish cars showed an RMSE of 2434 and R-squared 0.86.

Similarly to the Ridge model, registration year was still a relevant factor unlike mileage and engine types, brand and model were found important.

DATASET	METHOD	RMSE	Rsquared	FACTOR 1	FACTOR 2	FACTOR 3
AUSTRALIA	Random Forest	2682	-	year	odometre	4cyl 2.0L Petrol
	Multiple	3183	0.878	year	odometre	cab chassis
	Lasso	3197	0.877	6cyl 4.2L TurboDiesel	year	fj cruiser
	Ridge	3221	0.876	6cyl 4.2L TurboDiesel	year	8cyl 5.6L Petrol
	Decision Tree	5174	-	year	4cyl 3.0L TurboDiesel	hatch
GERMANY	Random Forest	1268	-	year	small car	kilometer
	Lasso	1636	0.733	7er	year	Z reihe
	Multiple	1638	0.732	year	kilometer	convertible
	Ridge	1648	0.731	year	A5	Lybra
	Decision Tree	2047	-	year	convertible	small car
SPAIN	Random Forest	1992	-	months old	BMW	Mercedes Benz
	Multiple	2434	0.864	months old	kms	Land Cruiser
	Lasso	2434	0.864	Land Cruiser	months old	Porsche
	Ridge	2447	0.863	Land Cruiser	months old	X6
	Decision Tree	4532	-	months old	gasoline	diesel

Fig.5 Summary of the results of the research

D. Decision Tree Regression

Decision Tree provides a visual, easy to interpret image of the data. They consist in a tree-shaped structure in which leaves represent decision rules and branches the features that lead to the decision rule. On the other hand, normally they do not provide great accuracy.

Running the model in the test samples, gave for the Australian used cars dataset an RMSE of 5174. For the German used cars the RMSE was 2047 and for the Spanish used cars, the RMSE was 4532.

Considering the factors, for all 3 datasets registration year was the most important factor. For Australian and German cars, the engine type and the vehicle type were found as well important. For the Spanish cars, it was found that whether the car used gasoline or diesel were the second important factor after registration year.

E. Random Forest Regression

It has been shown the Decision Tree model gave far worse results than the previous linear models; the alternative is to use Random Forest. This method uses the training data to train a chosen number of Decision Trees, in our research 150, out of the same amount of different data samples; this technique is called Bootstrap or Bagging. Then, it outputs a Decision Tree using the mean prediction of all the trees. One main advantage of Random Forest with respect to Decision Tree is that it prevents overfitting. On the other hand, as a general rule, it can be said that the more trees are implemented, the better results but this obviously will carry a computational cost.

This last model found these results. For Australia data, the RMSE was 2682. For the German cars, 1268. And for the Spanish cars the RMSE was 1992. It can be seen a huge result improvement compared to the Decision Tree.

The model showed as well that little improvement was found after the 50th tree as in the figure 6 example for the Australian cars (plots for the German and Spanish cars are practically the same as for the Australian cars).

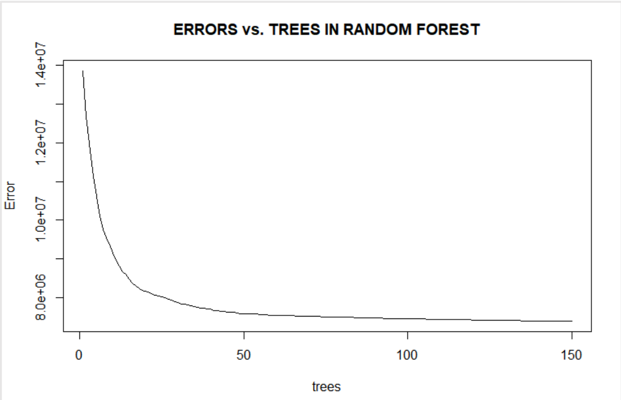

Fig.6 Random Forest model applied to German used cars data

V. CONCLUSIONS

When comparing the linear models used during the research, it was found that they provided a good fit for the data and that their performance was similar; in particular, the results for linear regression and Lasso were almost identical. It can be said that the dimensionality of the dataset was not high enough to benefit from using the Lasso model.

Meanwhile, the Ridge model showed a slightly worse performance. This method is usually used when there is multicollinearity between the variables, as this was not the case in the data used for the research, the method did not show a result improvement.

Decision tree model showed a bad performance with relation to the other models. This method is the less sophisticated of all and its use was more illustrative and used as an introduction to Random Forest.

Finally, the research found that Random Forest performed much better than the linear models. It was found for all datasets an important gap between its performance compared to the second-best model whether it was multiple or Lasso regression. The training stage was run for 150 trees and it was found that the model resulted in little improvement after the 50^{th} tree. Probably, if larger datasets were used, the training stage would have benefit of the use of more trees to build the model.

In terms of the factors that determine the price, it was found the registration year is in all cases in the first or second place; whereas mileage that initially could have been expected as well to be a top factor showed a variable importance depending on the methods and the datasets, although it was relevant for the best performing methods in Australian and German cars.

The research wants as well to emphasise that the results found are interpretable only from each dataset point of view. The aim of this research was not to compare the different markets: first because the data was scrapped from different data sources for different potential users, second because the variables used in each case different as well as the prices ranges.and, finally because probably the type of cars buyers are interested in differ for each country, particularly a big country like Australia as opposite of smaller European countries.

From a critical point of view, the use of the results of this research could be improved using larger datasets. As well, other techniques could be used as for example Gradient Boosting which could improve the performance of tree-based methods used in this research.

Another interesting aspect could have been to try to verify the performance of the models in the different datasets; in which case, only common variables should be used and some scaling transformation carried out to standardise the data.

VI. References

[1] E. Go, "We Made a Price Prediction Model for Used Cars in the Philippines," Medium, 17 08 2019. [Online]. Available: https://medium.com/@elysekatrina.go/price-prediction-model-for-2nd-hand-cars-f1801d8c8d47. [Accessed 03 2020].

[2] S. J. Cho, "The Determinants of Used Rental Car Prices," Journal of Economic Research, 23 08 2005. [Online]. Available: https://docplayer.net/12468945-The-determinants-of-used-rental-car-prices.html. [Accessed 02 2020].

[3] M. S. Richardson, "Determinants of Used Car Resale Value," The Colorado College, 2009.

[4] T. Andrews and C. Benzing, "The Determinants of Price in Internet Auctions of Used Cars," International Atlantic Economic Society, 2007.

[5] S. Kuiper, "Introduction to Multiple Regression: How Much Is Your Car Worth?," Journal of Statistics Education, 2008.

[6] J. Du, L. Xie and S. Schroeder, "Practice Prize Paper: PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution," Marketing Science, 2009.

[7] J.-D. Wu, C.-C. Hsu and H.-C. Chen, "An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference," ScienceDirect, 2009.

[8] M. Listiani, "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application," Hamburg University of Technology, 2009.

[9] S. Gongqi, W. Yansong and Z. Qiang, "New model for residual value prediction of used cars based on BP neural network and non-linear curve fit," IEEE Computer Society, 2011.

[10] S. Pudaruth, "Predicting the price of used cars using machine learning techniques," International Journal of Information & Computation Technology, 2014.

[11] N. H. C. S. P. Saamiyah Peerun, "Predicting the Price of Second-hand Cars using Artificial Neural Networks," Proceedings of the Second International Conference on Data Mining, Internet Computing, and Big Data, 2015.

[12] S. Lessmann and S. Voss, "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy," International Journal of Forecasting, 2017.

[13] K. Noor and S. Jan, "Vehicle Price Prediction System using Machine Learning Techniques," International Journal of Computer Applications, 2017.

[14] X. Zhang, Z. Zhang and C. Qiu, "Model of Predicting the Price Range of Used Car," 2017.

[15] N. Pal, A. P. K. P. D. Sundararaman and S. S. Palakurthy, "How much is my car worth? A methodology for," Future of Information and Communications Conference (FICC) , 2018.

[16] E. Gokce, "Predicting Used Car Prices with Machine Learning Techniques: Comparing Performance of Five Different ML Models," Towards Data Science, 01 2020. [Online]. Available: https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952. [Accessed 02 2020].

[17] Statista.com, "Transportation & Logistics: Vehicles & Road Traffic," Statista.com, 2020. [Online]. Available: https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/. [Accessed 03 2020].

[18] AnythingResearch, "2020 U.S. Industry Statistics & Market Research - Used Car Dealers," AnythingResearch, 03 2020. [Online]. Available: https://www.anythingresearch.com/industry/Used-Car-Dealers.htm. [Accessed 03 2020].

[19] Technavio, "Car Leasing Market by End-user and Geography - Forecast and Analysis 2020-2024," Infiniti Research Limited, 02 2020. [Online]. Available: https://www.technavio.com/report/car-leasing-market-industry-analysis. [Accessed 03 2020].

[20] S. Lessmann, M. Listiani and S. Voß, "DECISION SUPPORT IN CAR LEASING: A," ICIS 2010 Proceedings, 2010.