IDENTIFICATION OF CAR RESALE VALUE PREDICTION

1. INTRODUCTION

Predicting the price of used cars in both an important and interesting problem. According to data obtained from the National Transport Authority [1], the number of cars registered between 2003 and 2013 has witnessed a spectacular increase of 234%.From 68, 524 cars registered in 2003, this number has now reached 160, 701. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. It is reported in [2] that the sales of new cars has registered a decrease of 8% in 2013.In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party –usually a bank, insurance firm or other financial institutions) in which the buyer mustpay fixed instalments for a pre-defined number of months/years to the seller/financer. After the lease period is over, the buyer has the possibility to buy the car at its residualvalue, i.e. its expected resale value.seller/financers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financer at the beginning, the instalments will be higher for the clients who will certainly then opt for another seller/financer. If the residual value is over-estimated, the instalments will be lower for the clients but then the seller/financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of very high commercial importance as well. Manufacturers' from Germany made a loss of 1 billion Euros in their USA market because of mis-calculating the residual value of leased cars [3]. Most individuals in Mauritius who buy new cars are also very apprehensive about the resale value of their cars after certain number of years when they will possibly sell it in the used cars market. Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance. Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven. Other factors such as the type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), safety index, its size, number of doors, paint colour, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, its physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and otheroptions such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well. Some special factors which buyersattach importance in Mauritius is the local of previous owners, whether the car had been involved in serious accidents and whether it is a lady-driven car. The look andfeel of the car certainly contributes a lot to the price. As we can see, the price depends on a large number of factors. Unfortunately, information about all these factors are not always available and the buyer must make the decision to purchase at a certain price based on few factors only. In this work, we have considered only a small subsetof the factors mentioned above. More details are provided in Section III.This paper is organised as follows. In the next section, a review of related work is provided. Section III describes the methodology while in section IV, we describe, evaluate and compare different machine learning techniques to predict the price of used cars. Finally, we end the paper with a conclusion with some pointers towards future work.

ABSTRACT

Now-a-days, with the technological advancement, Techniques like Machine Learning, etc are being used on a large scale in many organisations. These models usually work with a set of predefined data-points available in the form of datasets. These datasets contain the past/previous information on a specific domain. Organising these datapoints before it is fed to the model is very important. This is where we use Data Analysis. If the data fed to the machine learning model is not well organised, it gives out false or undesired output. This can cause major losses to the organisation. Hence making use of proper data analysis is very important.

About Dataset:

The data that we are going to use in this example is about cars. Specifically containing various information datapoints about the used cars, like their price, color, etc. Here we need to understand that simply collecting data isn’t enough. Raw data isn’t useful. Here data analysis plays a vital role in unlocking the information that we require and to gain new insights into this raw data.

Consider this scenario, our friend, Otis, wants to sell his car. But he doesn’t know how much should he sell his car for! He wants to maximize the profit but he also wants it to be sold for a reasonable price for someone who would want to own it. So here, us, being a data scientist, we can help our friend Otis.

Let’s think like data scientists and clearly define some of his problems: For example, is there data on the prices of other cars and their characteristics? What features of cars affect their prices? Colour? Brand? Does horsepower also affect the selling price, or perhaps, something else?

As a data analyst or data scientist, these are some of the questions we can start thinking about. To answer these questions, we’re going to need some data. But this data is in raw form. Hence we need to analyze it first. The data is available in the form of `.csv/.data` format with us

Modules needed:

- **pandas:** Pandas is an opensource library that allows you to perform data manipulation in Python. Pandas provide an easy way to create, manipulate and wrangle the data.

- **numpy:** Numpy is the fundamental package for scientific computing with Python. `numpy` can be used as an efficient multi-dimensional container of generic data.
- **matplotlib:** Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of formats.
- **seaborn:** Seaborn is a Python data-visualization library that is based on matplotlib. Seaborn provides a high-level interface for drawing attractive and informative statistical graphics.
- **scipy:** Scipy is a Python-based ecosystem of open-source software for mathematics, science, and engineering.

Steps for installing these packages:

- If you are using anaconda- jupyter/ syder or any other third party softwares to write your python code, make sure to set the path to the "scripts folder" of that software in command prompt of your pc.
  - Then type – pip install package-name
    Example:

```
pip install numpy
```

  -
  - Then after the installation is done. (Make sure you are connected to the internet!!) Open your IDE, then import those packages. To import, type – import package name

```
import numpy
```

  -

Steps that are used in the following code (Short description):

- Import the packages
- Set the path to the data file(.csv file)
- Find if there are any null data or NaN data in our file. If any, remove them
- Perform various data cleaning and data visualisation operations on your data. These steps are illustrated beside each line of code in the form of comments for better understanding, as it would be better to see the code side by side than explaining it entirely here, would be meaningless.
- Obtain the result!

2. **Methodology**

Data was collected from <<petites announces>> found in daily newspapers such as L'Express [8] and Le Defi [9]. We made sure that all the data was collected in less than one month interval as time itself could have an appreciable impact on the price of
cars. In Mauritius, seasonal patterns is not really a problem as this does not really affect the purchase or selling of cars. The following data was collected for each car:

make, model, volume of cylinder (funnily this is usually considered same as horsepower in Mauritius), mileage in km, year of manufacture, paint colour, manual/automatic and price. Only cars which had their price listed were recorded.756 Sameerchand Pudaruth

Because many of the columns were sparse they were removed. Thus, paint colour and manual/automatic features were removed. The data was then further tweaked to remove records in which either the age (year) or the cylinder volume was not available. Model was also removed as it would have been extremely difficult to get enough records for all the variety of car models that exist. Although data for mileage was sparse, it was kept as it is considered to be a key factor in determining the price of used cars. A sample of the collected data is shown below in Table 1.

Table 1. Sample Data Collection

#	MAKE	CYLINDER VOLUME (CC)	YEAR	MILEAGE/KM	PRICE (RS)
1	TOYOTA	1300	2007	38000	410000
2	NISSAN	1500	2007	50000	325000
3	HONDA	1500	2005	59000	385000
4	TOYOTA	1000	2007	59000	360000
5	TOYOTA	1300	1989	62665	50000
6	TOYOTA	1500	2008	67000	615000
7	TOYOTA	1500	2008	69000	575000
8	TOYOTA	1490	2006	73000	450000
9	TOYOTA	1600	2006	82000	550000
10	TOYOTA	1000	2006	85000	325000
11	TOYOTA	1500	2000	113000	325000
12	TOYOTA	1500	2000	129000	218000
13	NISSAN	1500	2001	145000	195000

Initially, 400+ records were collected. However, after further pruning, for example, we kept only the three of the most popular makes in Mauritius, i.e. Toyota, Nissan and Honda. In particular, we removed all makes for which there were less than 10 records. Regarding the cylinder volume, for some cars, it was provided in a range. We then opted for the average value of the range. The final database contained only 97 records: Toyota (47), Nissan (38) and Honda (12). The values are then preprocessed in a form amenable to further processing using machine learning techniques. The minimum and maximum values for some numerical feature are shown in Table 2.

Table 2. Minimum and Maximum Values

#	CYLINDER VOLUME (CC)	YEAR	PRICE (RS)
Minimum	1000	1988	27, 000
Maximum	2160	2013	825, 000

4. Implementation and Evaluation

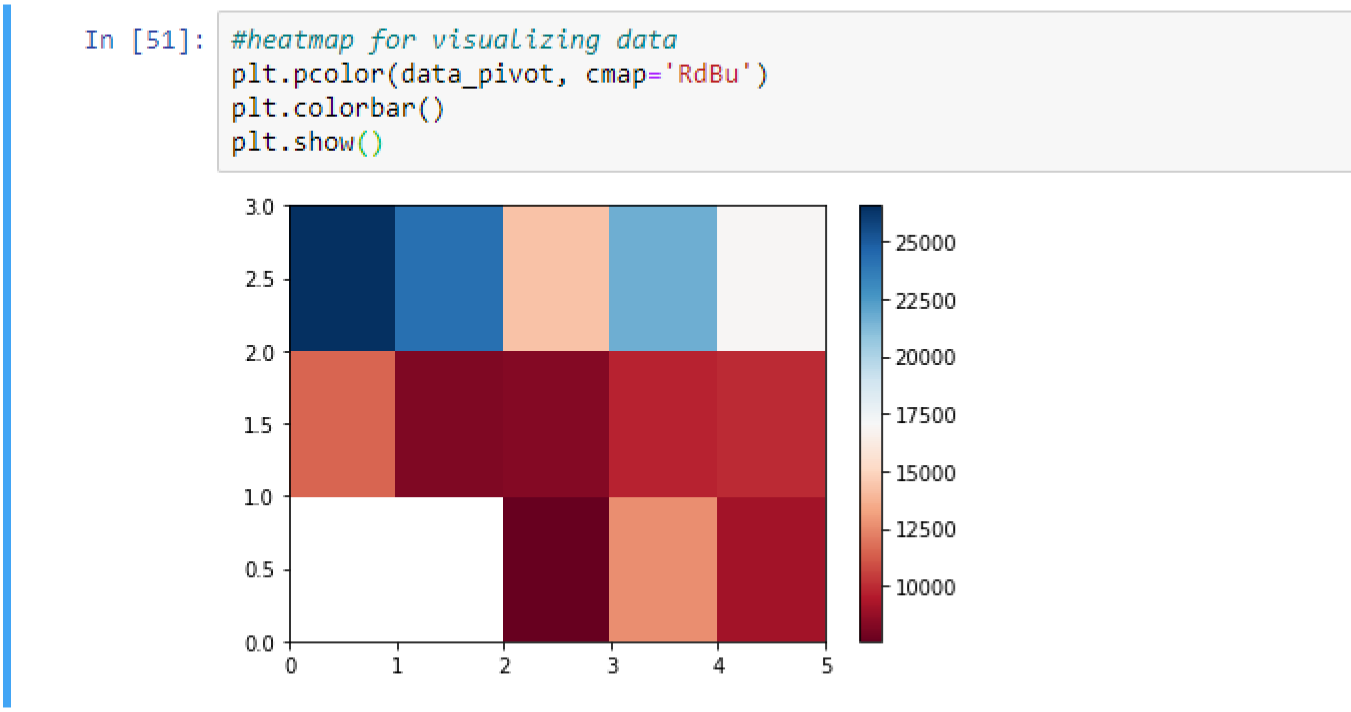
4.1. Multiple Linear Regression Analysis

The lack of mileage information for most of the cars did not allow us to use it to forecast the price. The Pearson correlation coefficient (r) was computed between different pairs of features [10]. The summarised results are shown below in Table 3

Table 3. Matrix of Pearson Correlation Coefficients

	Cylinder volume	Year	Mileage	Price
Cylinder volume	1	-	-	0.33
Year	-	1	-0.33	0.81
Mileage	-	-0.33	1	-0.35
Price	0.33	0.81	-0.35	1

The value of r was found to be -0.33 between year and mileage, which means that there is a weak negative correlation between the year in which the car was manufactured and its mileage. The relation is not strong enough to use year information to predict mileage information. The relation between mileage and price is also very similar. On the other hand, there is a weak positive correlation betweencylinder volume and price. This means that on average prices for cars with highercylinder volume tends to be slightly higher. As anticipated, there is a very highcorrelation between the price of a car and the year in which it was manufactured. Newcars have higher prices. Among the factors considered, we can see thatyear has the most impact. Since all the three makes that we have considered are Japanese manufacturers, we are assuming this will not impact on the price



In [51]:

```
#heatmap for visualizing data
plt.pcolor(data_pivot, cmap='RdBu')
plt.colorbar()
plt.show()
```

5. Evaluation and Conclusion

In this paper, four different machine learning techniques have been used to forecast the price of used cars in Mauritius. The mean error with linear regression was about Rs51, 000 while for kNN it was about Rs27, 000 for Nissan cars and about Rs45, 000 for Toyota cars. J48 and NaiveBayes accuracy dangled between 60-70% for different combinations of parameters. The main weakness of decision trees and naïve bayes is their inability to handle output classes with numeric values. Hence, the price attributehad to be classified into classes which contained a range of prices but this evidently introduced further grounds for inaccuracies. The main limitation of this study is the low number of records that have been used. As future work, we intend to collect more data and to use more advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices.

6. REFERENCES

[1] NATIONAL TRANSPORT AUTHORITY. 2014. Available from:

http://nta.gov.mu/English/Statistics/Pages/Archive.aspx [Accessed 15 January 2014].

[2] MOTORS MEGA. 2014. Available from: http://motors.mega.mu/news/2013/12/17/auto-market-8-decrease-sales-new-cars/ [Accessed 17 January 2014].

[3] LISTIANI, M., 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Thesis (MSc). Hamburg University of Technology.

[4] RICHARDSON, M., 2009. Determinants of Used Car Resale Value. Thesis (BSc). The Colorado College.

[5] WU, J. D., HSU, C. C. AND CHEN, H. C., 2009. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications. Vol. 36, Issue 4, pp. 7809-7817.

[6] DU, J., XIE, L. AND SCHROEDER S., 2009. Practice Prize Paper - PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, Vol. 28, Issue 4, pp. 637-644.

[7] GONGGI, S., 2011. New model for residual value prediction of used cars based on BP neural network and non-linear curve fit. In: Proceedings of the 3rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.

[8] LEXPRESS.MU ONLINE. 2014. Available from: http://www.lexpress.mu/ [Accessed 17 January 2014]

[9] LE DEFI MEDIA GROUP. 2014. Available from: http://www.defimedia.info/ [Accessed 17 January 2014]

[10] GELMAN, A. AND HILL, J., 2006. Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, New York, USA.