Resale value preditcion Using Watson Auto AI

By-Darpan Gupta

Contents

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 must pay 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 residual value, i.e. its expected resale value. Thus, it is of commercial interest to 754 Sameerchand Pudaruth 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

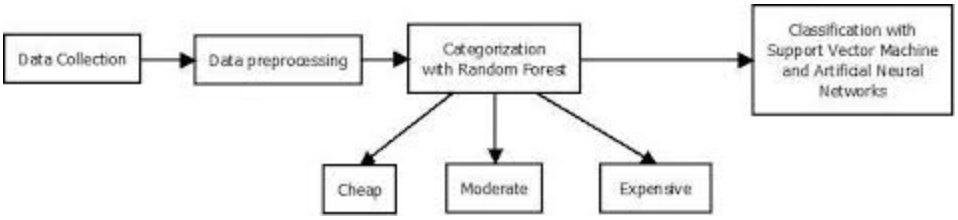
automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well. Some special factors which buyers attach importance in Mauritius is the local of previous owners, whether the car had been involved in serious accidents and whether it is a lady-driven car. The look and feel 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 subset of the factors mentioned above. More details are provided in Section III.

2. Literature Survey

Machine learning is a form of artificial intelligence which compose available computers with the efficiency to be trained without being veraciously programmed. Machine learning interest on the extensions of computer programs which is capable enough to modify when unprotected to new-fangled data. Machine learning algorithms are broadly classified into three divisions, namely; Supervised learning, Unsupervised learning and Reinforcement learning. Supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct answer. After that, machine is provided with new set of examples so that supervised learning algorithm analyses the training data and produces a correct outcome from labelled data. Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Unlike, supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabelled data by our-self.

3. Theoretical Anaslysis

A) Block Diagraam:-

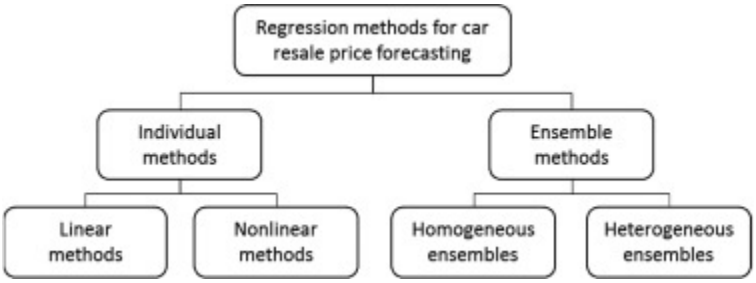


4. EXPERIMENTAL INVESTIGATIONS

Over the last two decades there have been a large number of empirical studies analysing land prices. Kilpatrick showed the usefulness of time-series regression model which used economic data to provide forecast of Central Business District (CBD) land price in moving market. Wilson et al studied the residential property market accounts for a substantial proportion of UK economic activity. Valuers estimate property values based on current bid prices. In this paper, the national housing transaction data was trained using Artificial Neural Networks (ANN), which forecasts future trend of the housing market. Mark and John developed a regression model with vacant land sales. The model explained up to 93% of the market values. Wang and Tian used the wavelet Neural Network (NN) to forecast the real estate price index. This kind of wavelet NN integrated the merit of the wavelet analysis and the tradition NN. It also compared the forecasting result with smoothing method and the NN forecast. Zhangming forecasted the real estate price index by using the Back Propagation (BP) NN. The BPN used the sigmoid function. Tinghao used the Auto Regressive Integrated Moving Average (ARIMA) model and carried the demonstrative analysis on year data from 1998 to 2006. He used the established model to make the forecast to the real estate price index of 2007. A hedonic regression on the price of land suggested that de facto policy differences between political jurisdictions have had a significant effect on land prices between1970 and 1980. Steven and Albert used 46,467 residential properties spanning 1999 - 2005 and demonstrated that using matched pairs that relative to linear hedonic pricing models, ANN generate lower dollar pricing errors, had greater pricing precision out-of-sample, and extrapolate better from more volatile pricing environments. ANN is better suited to hedonic models that utilize large numbers of variables. Sampath kumar and Santhi studied the land price trend of Sowcarpet which is the central part. They developed statistical model using economic factors and predicted that the annual rise in land price would be of 17%. Urmila reported that the past trends were analysed to

ascertain the rate of growth or decline and the trends are used in forecasting. Economic parameters might be introduced to formulate more realistic relationship. Some of the other techniques they Mansural Bhuiyan and Mohammad Al Hasan 2016 use is regression, deep learning to learn the nature of models from the previous results (the property/land which were sold off previously which are used as training data).There are different models used such as linear model data using only one feature, multivariate model, using several features as its input and polynomial model using the input as cubed or squared and hence calculated the root mean squared error (RMS value) for the model.

5. FLOWCHART



6. RESULT

RMSE			
S.no	Algorithms	RMSE Error	Accuracy
1	Random forset	0.012	90%
2	Neural Network	0.590	60%
3	Gradient Boosted	0.573	65%
4	Bagging	0.563	70%
5	Support Vector machine	0.636	58%
6	Multiple Regression	0.70	55%

7 . ADVANTAGES & DISADVANTAGES

Multiple Regression Technique - Regression analysis is widely used for forecasting. Regression analysis is used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. If more independent variables are added, it is able to determine an estimating equation

that describes the relationship with greater accuracy. Multiple regressions look at each independent variable and test whether it contributes significantly to the way the regression describes the data.
Advantages: The estimates of the unknown parameters obtained from linear least squares regression are the optimal. Estimates from a broad class of possible parameter estimates under the usual assumptions are used for process modelling. It uses data very efficiently. Good results can be obtained with relatively small data sets.

Disadvantages: The outputs of regression can lie outside of the range [0,1]. It has limitations in the shapes that linear models can assume over long ranges. The extrapolation properties will be possibly poor. It is very sensitive to outliers. It often gives optimal estimates of the unknown parameters.

Neural Network Technique – NN is a computational technology from the artificial intelligence discipline whose architecture emulates the network of nerve cells in the human brain. A NN is a parallel distributed information-processing structure consisting of processing elements (PEs) which contains local memory. NN architecture such as a standard BP NN can be developed by using the various indicators as PEs to be investigated upon. The approach presents the application of NN for modelling the land price trend with the support of economic and social factors. NN model is constructed with 13 indicators that are PEs with one bias node as input. All the input values are normalized using the MinMax.

Advantage: Neural Network lies in their ability to outperform nearly every other Machine Learning algorithms, but this goes along with some disadvantages that we will discuss and lay our focus on during this post. Like I already mentioned, to decide whether or not you should use Deep Learning depends mostly on the problem you are trying to solve with it. For example, in cancer detection, a high performance is crucial because the better the performance is the more people can be treated. But there are also Machine Learning problems where a traditional algorithm delivers a more than satisfying result.

Linear Regression - To establish baseline performance with a linear classifier, we used Linear Regression to model the price targets, Y, as a linear function of the data, X

Advantage: A linear model can include more than one predictor as long as the predictors are additive. the best fit line is the line with minimum error from all the points, it has high efficiency but sometimes this high efficiency created.
Disadvantage: Linear Regression Is Limited to Linear Relationships. Linear Regression

Only Looks at the Mean of the Dependent Variable. Linear Regression Is Sensitive to Outliers. Data Must Be Independent

Support Vector Regression -We used the linear SVR and also the polynomial and Gaussian kernels for regression of target prices. The linear SVR estimates a function by maximizing the number of deviations from the actually obtained targets Yn within the normalized margin stripe, while keeping the function as flat as possible. In other word, the magnitude of the error does not matter as long as they are less than, and flatness in this case means minimize w. For a data set of N target prices with M features, there are feature vectors Xn $\in$ RM where n = 1,...,N and the targets Yn corresponding to the price of real estate properties. The SVR algorithm is a convex minimization problem that finds the normal vector w $\in$ RM of the linear function as follows.

Advantages: It has a regularization parameter, which makes the user think about avoiding over-fitting. Secondly it uses the kernel trick, so you can build in expert knowledge about the problem via engineering the kernel. Thirdly an SVM is defined by a convex optimization problem (no local minima) for which there are efficient methods (e.g. SMO). Lastly, it is an approximation to a bound on the test error rate, and there is a substantial body of theory behind it which suggests it should be a good idea.

Disadvantages: Theory only really covers the determination of the parameters for a given value of the regularization and kernel parameters and choice of kernel. In a way the SVM moves the problem of overfitting from optimizing the parameters to model selection. Sadly, kernel models can be quite sensitive to overfitting the model selection criterion.

k-Nearest Neighbours (KNN) -k-Nearest-Neighbour (KNN) is a non-parametric instance-based learning method. In this case, training is not required. The first work on KNN was submitted by Fix & Hodges in 1951 for the United States Air-force. The algorithm begins by storing all the input feature vectors and outputs from our training set. For each unlabelled input feature vector, we find the k nearest neighbours from our training set. The notion of nearest uses Euclidean distance in the m-dimensional feature space. For two input vectors x and w

Advantages: The K-Nearest Neighbour (KNN) Classifier is a very simple classifier that works well on basic recognition problems.

Disadvantage: KNN algorithm is that it is a lazy learner, i.e. it does not learn anything from the training data and simply uses the training data itself for classification. To

predict the label of a new instance the KNN algorithm will find the K closest neighbours to the new instance from the training data, the predicted class label will then be set as the most common label among the K closest neighbouring points. The main disadvantage of this approach is that the algorithm must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training examples. Another disadvantage of this approach is that the algorithm does not learn anything from the training data, which can result in the algorithm not generalizing well and also not being robust to noisy data. Further, changing K can change the resulting predicted class label.

Random Forest Regression -The Random Forest Regression (RFR) is an ensemble algorithm that combines multiple Regression Trees (RTs). Each RT is trained using a random subset of the features, and the output is the average of the individual RTs. The sum of squared errors for a tree T

Advantages: There is no need for feature normalization. Individual decision trees can be trained in parallel. Random forests are widely used. They reduce overfitting.

Disadvantages: They’re not easily interpretable. They’re not a state-of-the-art

8. APPLICATIONS

Machine Learning is applied at Netflix and Amazon as well as for Facebook's face recognition. For you as a user, Machine Learning is for example reflected in the possibility of tagging people on uploaded images. In fact, Facebook has the largest face database in the world. The data fed by users into the social network is used by Facebook to optimize and train Machine Learning systems in terms of visual recognition.

Another application of Machine Learning that is now firmly integrated into everyday life is the automatic detection of spam that is integrated into almost all e-mail programs. Within the scope of spam detection, the data contained in the e-mails is analysed and categorised. The "spam" and "non-spam" patterns are used in this respect. If an e-mail is recognized as junk mail, the program learns to identify spam mails even more efficiently. Other areas of application for Machine Learning include search engine ranking, combating cybercrime and preventing computer attacks.

9&10. Conclusion and Future Work

The main goal of this project is to determine the prediction for prices which we have successfully done using different machine learning algorithms like a Random forest, multiple regression, Support vector machine, gradient boosted trees, neural networks, and bagging, so it's clear that the random forest have more accuracy in prediction when compared to the others and also my research provides to find the attributes contribution in prediction. So I would believe this research will be helpful for both the peoples and governments and the future works are stated below Every system and new software technology can help in the future to predict the prices. price prediction this can be improved by adding many attributes like surroundings, marketplaces and many other related variables to the houses. The predicted data can be stored in the databases and an app can be created for the people so they would have a brief idea and they would invest the money in a safer way. If there is a possibility of realtime data the data can be connected to the H2O and the machine learning algorithms can be directly connected with the interlink and the application environment can be created.

11. BIBILOGRAPHY

- Bhagat, N., Mohokar, A. and Mane, S. (2016). House price forecasting using data mining, International Journal of Computer Applications 152(2): 23–26. URL: http://www.ijcaonline.org/archives/volume152/number2/26292-2016911775
- Gu, J., Zhu, M. and Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine, Expert Systems with Applications 38(4): 3383–3386.
- Li, L. and Chu, K.-H. (2017). Prediction of real estate price variation based on economic parameters, Applied System Innovation (ICASI), 2017 International Conference on, IEEE, pp. 87–90.
- Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data, Expert Systems with Applications 42(6): 2928–2934.
- Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques, Expert Systems with Applications 42(1): 259–268.
- Wu, L. and Brynjolfsson, E. (2009). The future of prediction: How google searches foreshadow housing prices and sales.
- Selim, H. (2009). Determinants of house prices in turkey: Hedonic regression versus artificial neural network, Expert Systems with Applications 36(2): 2843–2852.

12 APPENDIX

UI output Screenshot.