Used Cars Price Prediction using Supervised Learning Techniques

Pattabiraman Venkatasubbu, Mukkesh Ganesh

Abstract: ***The production of cars has been steadily increasing in the past decade, with over 70 million passenger cars being produced in the year 2016. This has given rise to the used car market, which on its own has become a booming industry. The recent advent of online portals has facilitated the need for both the customer and the seller to be better informedabout the trends and patterns that determine the value of a used car in the market. Using Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees, we will try to develop a statistical model which will beable to predict the price of a used car, based on previous consumer data and a given set of features. We will also be comparingthe prediction accuracy of these models to determine the optimal one.***

Keywords: ANOVA, Lasso Regression, Regression Tree, Tukey's Test

I. I. I. INTRODUCTION

The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. The emergence of online portals such as CarDheko, Quikr, Carwale, Cars24, and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the valueof the used car in the market. Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features.

Different websites have different algorithms to generate the retail price of the used cars, and hence there isn't a unified algorithm for determining the price. By training statistical models for predicting the prices, one can easilyget a rough estimate of the price without actually entering the details into the desired website. The main objective ofthis paper is to use three different prediction models to predict the retail price of a used car and compare their levelsof accuracy.

The data set used for the prediction models was created by Shonda Kuiper[1]. The data was collected from the 2005 Central Edition of the Kelly Blue Book and has 804 records of 2005 GM cars, whose retail prices have beencalculated. The data set primarily comprises of categorical attributes along with two quantitative attributes.

The following are the variables used:

Revised Manuscript Received on December 02, 2019

***** Correspondence Author

Pattabiraman Venkatasubbu*, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India. Email: pattabiraman.v@vit.ac.in

Mukkesh Ganesh, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India. Email: g.mukkesh2017@vitstudent.ac.in

Price: The calculated retail price of GM cars. The cars which were selected for this data set were all less than ayear old and were considered to be in good condition.
Mileage: The total number of miles the car has been driven
Make: The manufacturer of the car **Model:** The specific models for each car **Trim:**
Thetype of car model
Type: The car's body type
Cylinder: The number of cylinders present in the engine
Liter: The fuel capacity of the engine
Doors: The number of doors in the car
cruise: A categorical variable (binary), which represents whether cruise control is present in the car (coded 1 ifpresent) **sound:** A categorical variable (binary), that represents whether upgraded speakers are present in the car (coded 1 if present)
Leather: A categorical variable (binary), that represents whether the car has leather interiors (coded 1 ifpresent)

Using these attributes, we will try to predict the price by using the Statistical Analysis System (SAS) forexploratory data analysis.

II. LITERATURE SURVEY

Overfitting and underfitting come into picture when we create our statistical models. The models might betoo biased to the training data and might not perform well on the test data set. This is called overfitting. Likewise, the models might not take into consideration all the variance present in the population and performpoorly on a test data set. This is called underfitting. A perfect balance needs to be achieved between these two,which leads to the concept of Bias-Variance tradeoff. Pierre Geurts [2] has introduced and explained how bias-variance tradeoff is achieved in both regression and classification. The selection of variables/attribute plays a vital role in influencing both the bias and variance of the statistical model. Robert Tibshirani [3] proposed a new method called Lasso, which minimizes the residual sum of squares. This returns a subset of attributes which need to be included in multiple regression to get the minimal error rate. Similarly, decision trees suffer from overfitting if they are not pruned/shrunk. Trevor Hastie and Daryl Pregibon [4] have explained the concept of pruning in their research paper. Moreover, hypothesis testing using ANOVA is needed to verify whether the different groups of errors really differ from each other. This is explained by TKKim and Tae Kyun in their paper [5]. A Post-Hoc test needs to be performed along with ANOVA if the number of groups exceeds two.

Tukey's Test has been explored by Haynes W. in his research paper [6]. Using these techniques, we will create,train and test the effectiveness of our statistical models.

III. PROPOSED MODEL

A. Null Hypothesis

Even though the magnitude of overfitting has been reduced, Regression trees still suffer from overfitting even afterPruning. This leads to our following hypothesis.

Hypothesis: Multiple and Lasso Regressions are better at predicting price than the Regression Tree.

B. Training and Testing Data

The data is split into training(70% - 563 records) and testing(30% - 241 records) data sets through random sampling(seed was set to 2786).

C. Lasso Regression

Using Lasso regression on the training data set, we first select the subset of attributes which lead to optimal/least sum of squared error while predicting the price. It makes use of 10-fold cross-validation to "lasso" the optimal subset of attributes. It uses L1 regularization.

IV. CONCLUSION AND FUTURE ENHANCEMENT

The prediction error rate of all the models was well under the accepted 5% of error. But, on further analysis, the mean error of the regression tree model was found to be more than the mean error rate of the multiple regression andlasso regression models. Even though for some seeds the regression tree has better accuracy, its error rates are higherfor the rest. This has been confirmed by performing an ANOVA. Also, the post-hoc test revealed that the error rates in multiple regression models and lasso regression models aren't significantly different from each other. To get even more accurate models, we can also choose more advanced machine learning algorithms such as random forests, an ensemble learning algorithm which creates multiple decision/regression trees, which brings down overfitting massively or Boosting, which tries to bias the overall model by weighing in the favor of good performers. More data from newer websites and different countries can also be scraped and this data can be used to retrain these models to check for reproducibility.

REFERENCES

[1] Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[2] Geurts P. (2009) Bias vs Variance Decomposition for Regression and Classification. In: Maimon O., Rokach L. (eds) Data Miningand Knowledge Discovery Handbook. Springer, Boston, MA

] [3] Robert T. (1996) Regression Shrinkage and Selection Via the Lasso. In: Journal of the Royal Statistical Society: Series B(Methodological) Volume 58, Issue

[4] Hastie, Trevor, and Daryl Pregibon. Shrinking trees. AT & T Bell Laboratories, 1990.

Kim, Tae Kyun. "Understanding one-way ANOVA using conceptual figures." Korean journal of anesthesiology 70.1 (2017): 22.

[6] Haynes W. (2013) Tukey's Test. In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology. Springer,

New York, NY

[7] Jaccard, James, Michael A. Becker, and Gregory Wood. "Pairwise multiple comparison procedures: A review." Psychological Bulletin 96.3 (1984): 589.

[8] Dupac, Václac, ed. Sampling from a finite population. Marcel Dekker, Incorporated, 1981.

AUTHORS PROFILE

Pattabiraman Venkatasubbu obtained his Ph.D. from Bharathiar University, India. He has a total Professional experience of 19 years working in various prestigious institutions. He is currently a Professor at Vellore Institute ofTechnology, Chennai Campus, India. He has authored several books in the field of Computer Science. He is a Senior member of International Association of Computer Science and Information Technology (IACSIT) also he is member in various professional societies namely ACM, IEEE, ISTE, CSI, Society for Research in Information Security and Privacy- SRISP and Academy & Industry Research Collaboration Center (AIRCC). Dr. Pattabiraman's teaching and research expertise covers a wide range of subject area including Knowledge discovery and Data mining, Big Data Analytics, Machine Learning, Deep Learning, Database technologies,Data Structures and Analysis of Algorithms etc., He has also received several awards in his career.

Mukkesh Ganesh is a B.tech Computer Science student at Vellore Institute of Technology, Chennai. A budding ML and AI enthusiast, who is working on leveraging the power of AI to solve a variety of problems. He is currently researching on the detection of anomalies using deep learning methods and time series forecasting of kidnapping rates in India. He is also interested in the prospect of enhancing Edge Computing in the field of IOT, to decentralize the heavy server-level processing. His other areas of interest include algorithm design, parallel and distributed computing and theory of computation. He is also working on several projects which span across different fields in Computer Science. He is a finalist of Smart India Hackathon 2019 software edition, and was one of the youngest team leaders in the government organized compe