TABLE OF CONTENTS

LIST OF FIGURES

ABSTARCT:

The paper is concerned with statistical models to forecast resale prices of used cars. An empirical study is performed to explore how different degrees of freedom in the modeling process contribute toward forecast accuracy. First, a comparative analysis of alternative prediction methods evidences that random forest regression is particularly effective in resale price forecasting. It is also shown that the use of linear regression, the prevailing method in previous work, should be avoided. Second, empirical results evidence the presence of heterogeneity in resale price forecasting and identify methods that can automatically overcome its detrimental effect on forecast accuracy.

Finally, the study confirms that the sellers of used cars possess informational advantages over market research agencies, which enable them to forecast resale prices more accurately. This implies that sellers have an incentive to invest into an in-house forecasting solution, instead of basing pricing decisions on externally generated residual value estimates.

Chapter 1

INTRODUCTION:

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 installments for a pre-defined number of months/years to the seller/financier. 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 seller/financiers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financier at the beginning, the installments will be higher for the clients who will certainly then opt for another seller/financier. If the residual value is over-estimated, the installments will be lower for the clients but then the seller/financier may have much difficulty at selling these high-priced used cars at this over-estimated residual value 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. 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.

Chapter 2

Literature review

2.1 Car Sales Prediction Using Machine Learning Algorithms*Abstract:*

Abstract:

Sales prediction is the current numero trend in which all the business companies thrive and it also aids the organization or concern in determining the future goals for it and its plan and procedure to achieve it. The data about car sales are derived from various sources .sales of cars does not contain any independent variable since various factors such as horse power; model, width, fuel type, height, price, city-mileage, highway-mileage and manufacturer are the various features that influence the sales. In car sales prediction we first implement the methodology of analytic hierarchy process in order to get varied idea about how well the various criteria's in our dataset works and after this we apply the machine learning algorithms such as Linear regression, Random tree to get the best clusters and we process them in to random forest to get best accurate feature out of it. which is ultimately followed by Technique for Order of preference by similarity to ideal solution (TOPSIS) an tool which helps the researcher to arrive at a verdict when he/she faces the one or more pattern selection problem and the final resultant derived from all these methods gives the fittest feature which influences the customer in purchasing the car which indirectly gives the company or the research market a result in predicting the future sales for cars. Hence this paper not only provides its users with some stats, it also serves an guiding guardian by providing accurate results for purchasing a car.

INTRODUCTION

Every human being has got his/her business in this world .so the term business comes with certain tagged words such as sales, profit and loss. A concerns success is determined by its sales and the performance of its product in the industry is also identified by its sales. Hence in order to improve the standard of the firm /concern the strategies, the techniques the methodologies are built .in this process of development forecasting of certain key terms such as profit, loss, return of interest may pave the way for the successful venture of the concern yet when we look deep in to the details all

these key terms are conveniently related to the term called sales .sales prediction is one of the master trades of business which may open the gateways for obtaining knowledge about the existing market trends and the ways to conquer the market .planning is the first step in every activity we perform and hence knowing what lies ahead in terms of sales hugely aids the concern/organization in this planning process . sales prediction can be a massive support in order to perform tracking, cashflow and purchasing .sales forecasting provides the business minds an idea of about how much to buy and how much not to buy ,what are the risk can be taken in terms of revenue ,how to plan budget ,market trends, introduction of new products according to the organizations capability and ability ,what changes may happen if the plan fails are the areas where it helps to develop our ideas ,let us shift our focus from the generic business term to its detailed chunk of streams such as education,medical,automobile and lots of other streams. Once great interest in recent days lies in automobile industry. The first boon to the automobile industry came only after the industrialization renaissance period. Now the automobile industry creates, packs, runs, moves the world on its wheels and also allows other fields to flourish parallel along with them .yet it also has its own defaults since it is not cost friendly ,any single default can affect the whole system and it cause a huge failure in the market for its brand ,there is a huge amount of revenue being spend and huge amount of chunks being produced and exploited on daily basis hence sales forecasting in this industry can lead to the organized pattern of selling,buying,producing goods and even the taxes to be implemented also comes in to role .forecasting these sales in automobile industries can be performed with various and variety of technologies and one among them is the machine learning technique. That may help in classifying the prediction of an automobile for a say lets us take it as car the yearly sales of it if know before then it will provide the manufacturer the huge boost in designing it, getting spare parts, getting key parts and reducing the waste products and tracking its revenue model its generation and various other activity. The classifiers used such as logistic regression, decision tree and random forest provides us with accurate prediction results.

Literature Review

Machine learning models and bankruptcy prediction is a paper work which talks about the improvement that takes place in academics industry with the aid of machine learning algorithms in predicting bankruptcy. The data is derived from integrated resource of Salomon center database which contains the details about the North American firms from the period between 1985 to 2013 . This paper implements the usage of algorithms such as bagging, boosting, random forest and support vector machine for predicting bankruptcy even before the event occurs and a greater span

of comparative study takes place with the performance of these results with the results of logistic regression and neural networks. Original Altman's Z-score variables are used as predictive variables with addition of extra variables such as the operating margin,sales,growth measures related to assets, change in return-on-equity, change in price-to-book, and number of employees based on carton and Hofer(2006). And a comparison is made between the models and these variables, the machine learning techniques and the algorithm with most accuracy is determined. Handling class imbalance in customer churn prediction by j.Burez and D. van den poel suggests the customer the various ways to handle class imbalance in churn prediction. AUC and lift are the evaluation metrics with which the sampling methods are interrogated. The modeling techniques such as weighted random forest, gradient boosting are compared with other techniques. The better evaluation metrics and the best modelling techniques are found out with the help of each techniques accuracy and from past studies. Calling communities analysis and identification using machine learning techniques is the work that determines the worth of a particular customer with respect to his/her general pattern trait of the community that he/she hails from. The customer calling impressions can be told beforehand by making use of a classifier model and cluster analysis for detail selection. The attributes such as accuracy and computational performance are taken in to consideration for comparison of various machine learning techniques. Customer churn prediction using improved balanced random forests is the paper that explains the real time working model that had been used in china. Improved balanced random forest is the hybrid version of balanced random forest and weighted random forest, two interval variables had been introduced such as e and f where e is the middle point and f is the length of interval .Random distribution of these classes are maintained with the help of these variables .Hence it produces more accurate results than its other counterparts. A sampling based sentiment mining approach for e-commerce applications paper puts the limelight on how the customers are being influenced by the Online reviews which is a part of marketing strategy of the e-commerce platforms. Hence this issue is attempted with the help of mining techniques. The two sampling methods are also used for classification of imbalanced data. A modified support vector machine based ensemble algorithm is the methodology used by the researches to identify the performance prediction [10]. On the differential benchmarking of promotional efficiency with machine learning modelling (II): Practical applications presents two different databases of different categories such as non-seasonal and heavy seasonal and models are analyzed here .The detailed performance of four famous machine learning techniques that has huge complexity is been dissected in this work. Certain features of various machine learning algorithms do not perform well because of these databases. In order to gain more accurate dissection results and feature extraction there is a need to implement certain correct procedures that may influence the specificity of the behavior of certain

categories and product ranges. Linguistic features for review helpfulness prediction by Srikumar Krishnamurthy analyses what makes an Online review with the help of a predictive model .This model follows the methodology of extracting linguistic category features such as adjective feature, state verb feature and action verb features it also takes in to account the readability related features for prediction. Hence the hybrid set of features that are obtained after the analysis on two real-life review datasets gives the researches the best accuracy rate of all time. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data paper puts the focus on how this work can contribute to the real estate industry that may cause an adverse effect on the US housing market.an eight step methodologies are used for the dataset of 5359 townhouses in Fairfax County, Virginia. The dataset has been segregated in to training, validation and testing set then training parameters such as C4.5, RIPPER, Naïve Bayesian, and Adaboost are set and the model is trained and evaluated using training set and validation set respectively and this process is iterated until it gains an optimal error in training, validation and testing.Finaly these results are compared to gain the optimal accuracy results. Explaining machine learning models in sales prediction is a generic manuscript that discusses about the recent trends of predictive models, real time scenarios in order to gain a deep insight about buyers and seller's interaction and the forecasting of sales. Early churn prediction with personalized targeting in mobile social games is a manuscript that explains Customer churn .churn is defined by the act of a customer leaving a product for good. This churn are reduced to a greater extent by following the procedure of mapping the feature with the interest of the customer and pushing the notifications in order to drag back the customer in to the game .this manuscript implements the methodologies such as logistic regression for the simple object linear model ,decision trees for extracting redundancy from features random forest to be used in various situations .Naive Bayes for generating the models and gradient boosting for its popularity. Distributed customer behavior prediction using multiplex data: A collaborative MK-SVM approach is a paperwork that explains the challenges that are encountered by the traditional method in predicting customer behavior. It contains a huge datasets which are categorized as static, symbolic sequential, textual data and time series in its database collaborative multiple kernel support vector machine (C-MK-SVM) is the new technique that is used for distributed customer behavior prediction with the aid of multiplex data. Various sub models in this technique is used for global optimization. The results obtained through Computation tell the researches that it is best suited for customer behavior prediction performance and for its maximum computational speed. Customer churn prediction using improved balanced random forest is also one of the works on churn prediction.it undergoes the disadvantage of imbalanced in the data distribution. This improved random balanced forest also uses some other sampling techniques with it. IBRF's features are mis

classified minority class with higher penalties are iteratively learned by altering the class distributions. It works with the real time data such as bank customer database. When compared with other methodologies such as artificial neural networks, decision trees, and class-weighted core support vector machines (CWC-SVM) IBRF has more accurate prediction features.The process /model of the system is kicked off to action by collecting the data for car sales prediction from renowned data repositories and available databases around the world and then these data are preprocessed, then all these data sets are selected for training and the best dataset are classified and these classifiers are further trained using various methodology and the results are predicted and its performance are evaluated and finally the results are being displayed.

IMPLEMENTATION

AHP-The Analytic hierarchy process is a Boon to the data industry since it helps people in taking the complex decisions .any decision depends on at what point of time the decision is taken and how well it helps in growth of the product. Hence this AHP helps in taking not just one equitable decision but it stocks up the customer with the most desirable decision solution that can help them in achieving their objective to their problem. It can be implemented in various sectors where we can expect a huge amount of data .For example in automobile industry people can have a problem related to purchasing an automobile in this case AHP comes in to action where out of various types of automobiles it gives the customer a detailed report of desirable solutions. When we look deep down in to AHP process certain steps needs to be followed. We take the Automobile dataset and the attributes of the dataset are given a grading based on the weights of those attributes and in the next process these weighted attributes are been assessed based on its values .The points are pitted against each other in pairs and it is checked whether they can achieve its objective .The number of pairs to be drawn against each other are first entered and then these pairs are pitted against AHP priorities and each attributes are given parallel values in the integer range of one to nine where one represents significant, three represents moderate significance ,five represents strong significance, seven represents very strong significance ,nine represents .in our automobile dataset eight attributes such as city-mileage, highway-mileage, model name,height,width,price,fuel type,horsepower in which price serves with the value of high significance and the difference between each of these eight values are pitted against each other to gain the consistency rate .if that value is less than ten percent it is said to be of with good consistency After the obtaining the suggestion for the most likely decision

like price should have the highest priority the machine learning algorithms are applied to this dataset to obtain a better predictive mode.

Linear Regression:

The interaction that happens between any two fixed components which are also coined as variables and the method used to define or calculate the weightage of their bond is called regression analysis Their main focus is to make the novice understand the process that happens when there is a drastic change and how these changes happen due to the modification of predictors and the adverse effect it has on criterion variable .at the preliminary level the data generation serves as a key in order to understand this model. The always three vital parts needed are the criterion variable, the predictors and the disguised parameters. We also categorize them in to two simple and multiple (more than one variable).The tools or the metrics to measure the linear variable varies within these two types an unstoppable scale and categorical one .to put in precise terms correlation is what the linear regression is made of whereas the difference lies in their ability to distinct things.