Car Resale Value Prediction
Literature Survey

Various methodoligies used are

1.Used Car Price Prediction using K-Nearest Neighbor Based Model

Sameerchand Pudaruth[1] proposed predicting the Price of Used Cars using Machine Learning Techniques. In this paper, they collected the historical data of used cars in Mauritius from the newspapers and applied different machine learning techniques like decision tree, K-nearest neighbours, Multiple Linear Regression and Naïve Bayes algorithms to predict the price. This model has the mean error about Rs.27000 for Nissan cars and about Rs45000 for Toyota cars using KNN and around Rs51000 using linear regression. The accuracy of decision trees and NaïveBayes algorithm dangled between 60 to 70 percentile with different parameters and the overall training accuracy of the model is 61%. Nitis Monburinon et al. [2] proposed a prediction of Prices for Used Car by Using Regression Models. In this paper, the authors selected the data from the German ecommerce site. The main goal of this work is to find a suitable predictive model to predict the used cars price. They used different machine learning techniques for comparison and used the mean absolute error(MAE) as the metric. They proposed that their model with gradient boosted regression has a lower error with MAE value 0.28 and this gives the higher performance where linear regression has the MAE value 0.55, random forest with MAE value 0.35. Enis Gegic et al. [3] proposed Car Price Prediction using Machine Learning Techniques. In this paper, they proposed an ensemble model by collecting different types of machine learning techniques like Support Vector Machine, Random Forest and Artificial neural network. They collected the data from the web portal www.autopijaca.ba and build this model to predict the price of used cars in Herzegovina and Bosnia. The

accuracy of their model is 87%. Kanwal Noor and Sadaqat Jan[4] proposed Vehicle Price Prediction System using Machine Learning Techniques. In this paper, they proposed a model to predict the price of the cars through multiple linear regression method. They selected the most influencing feature and removed the rest by performing feature selection technique. The Proposed model achieved the prediction precision of about 98%. In this paper, a machine learning model is proposed to estimate the cost of the used cars using the K-Nearest Neighbor algorithm. The model is trained with used cars data for different trained and test ratios. Then the proposed model is cross-validated using K fold method to examine the performance to avoid the over fit.

2.Used Car Price Prediction Using Machine Learning Algorithm

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 [9]. 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 [6]. 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 [1]. 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 [2]. 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 [11]. 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 [8]. 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 [7]. 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 [5]. 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 [4]. 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 [3].

3.**Used Cars Price Prediction using Supervised Learning Technique**

Overfitting and underfitting come into picture when we create our statistical models. The models might be too 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 perform poorly on a test dataset. 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 iseeded to verify whether the different groups of errors really differ from each other. This is explained by TK Kim 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.