CAR RESALE VALUE PREDICTION

Introduction

With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright. 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 sellers/financers to be able to predict the salvage value (residual value) of cars with accuracy.

In order to predict the resale value of the car, we proposed an intelligent, flexible, and effective system that is based on using regression algorithms. Considering the main factors which would affect the resale value of a vehicle a regression model is to be built that would give the nearest resale value of the vehicle. We will be using various regression algorithms and algorithms with the best accuracy will be taken as a solution, then it will be integrated to the web-based application where the user is notified with the status of his product.

Literature Review

(1) Linear Regression:

It is an AI calculation dependent on administered learning. It plays out a relapse task. It is utilized to assess genuine qualities (cost of houses, number of calls, absolute deals and so forth) in view of nonstop variable(s). Here, we set up connections among free and ward factors by fitting a best line. This best fit line is known as relapse line and spoke to by a straight condition Y= a *X + b. Prior to understanding what direct relapse is, let us get ourselves acclimated with relapse. Relapse is a strategy for demonstrating an objective worth dependent on free indicators. This strategy is generally utilized for spreading and discovering circumstances and logical results connection between factors. Relapse methods generally vary depending on the quantity of autonomous factors.

In fig(1) it is very clear how the linear regression algorithm will work.

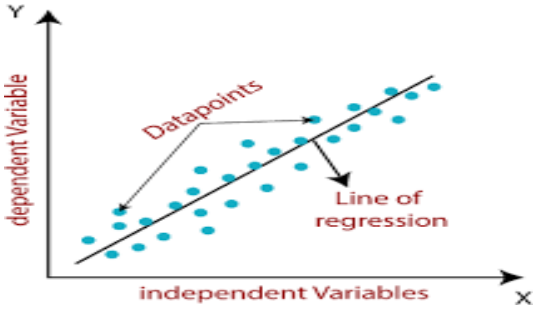
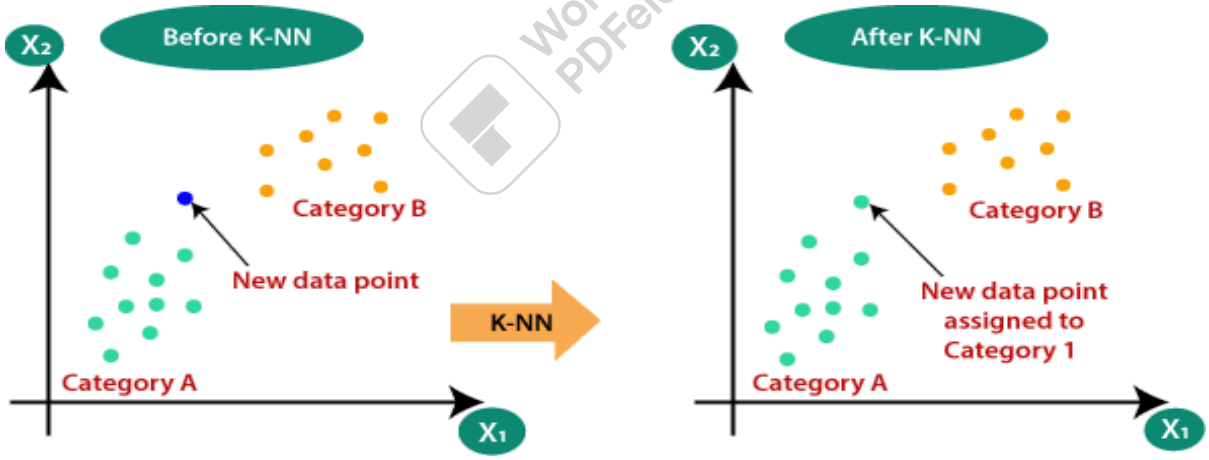


Fig (1)

(2) K-Nearest Neighbors

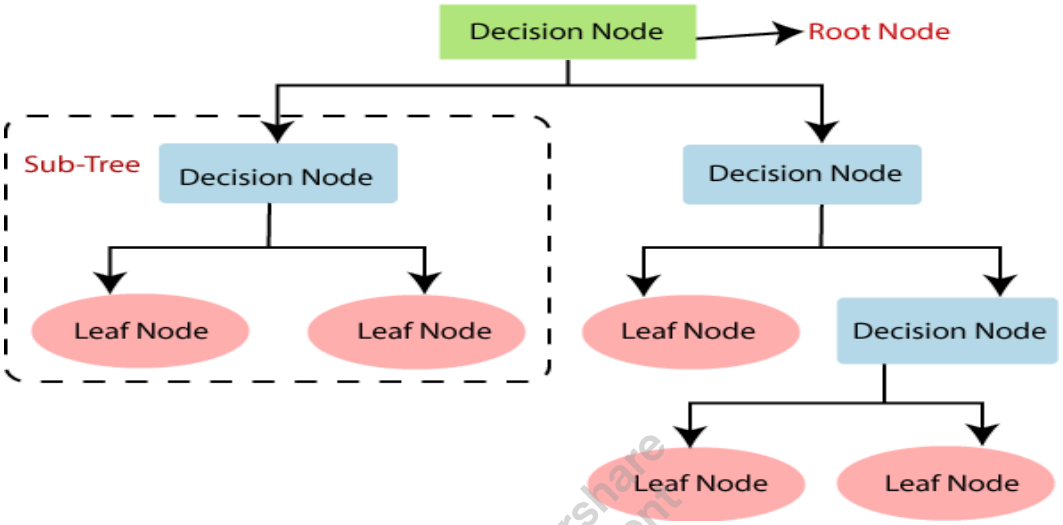
K-nearest neighbor is a machine learning technique in which the new (unknown) data is compared to all the existing records in order to locate the best match(es).Despite its apparent simplicity, a lot of take has to be taken in pre-processing the data otherwise we can easily go off-track. Only three attributes were considered namely the make, year and cylinder volume. The K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.



(3) Decision Tree

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes and leaf nodes. Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels. Whether or not all data points are classified as homogeneous sets is largely dependent on the complexity of the decision tree.

Smaller trees are more easily able to attain pure leaf nodes—i.e. data points in a single class. However, as a tree grows in size, it becomes increasingly difficult to maintain this purity, and it usually results in too little data falling within a given subtree. When this occurs, it is known as data fragmentation, and it can often lead to overfitting. As a result, decision trees have preference for small trees, which is consistent with the principle of parsimony in Occam's Razor; that is, "entities should not be multiplied beyond necessity." Said differently, decision trees should add complexity only if necessary, as the simplest explanation is often the best. To reduce complexity and prevent over-fitting, pruning is usually employed; this is a process, which removes branches that split on features with low importance.

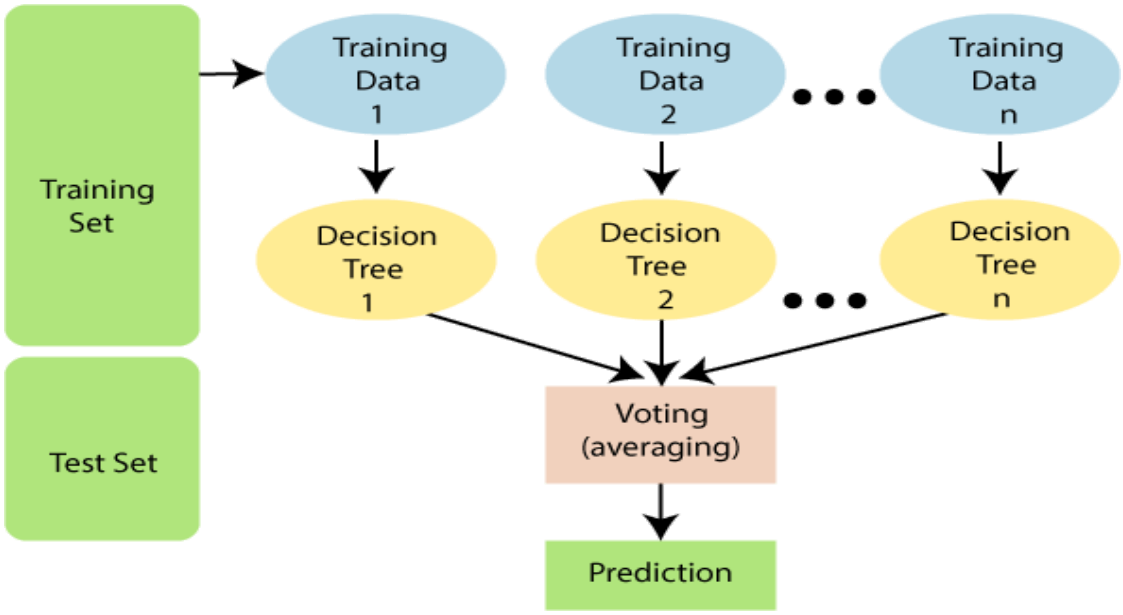


(4) Random forest classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.* As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.* As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The below diagram explains the working of the Random Forest algorithm:

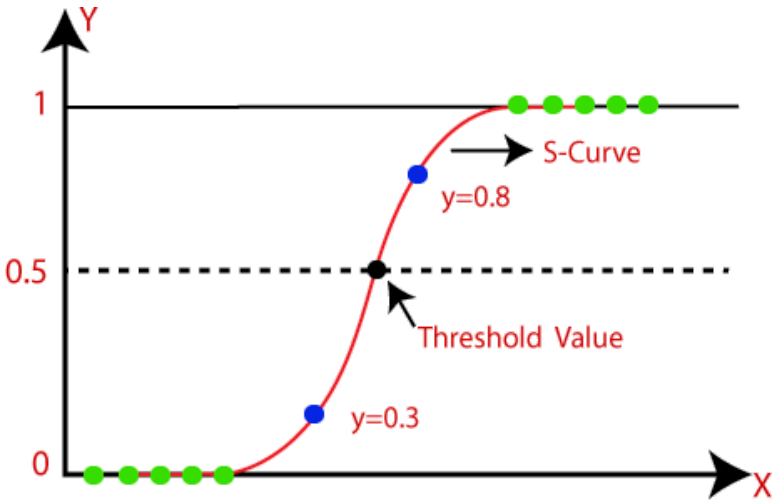


(5) Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.

Logistic Regression is much similar to Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

The below image is showing the logistic function:



REFERENCES:

1. Pudaruth, S., 2014. “Predicting the Price of Used Cars using Machine Learning Techniques.” Vol 4, Number 7 (2014), pp. 753-76.

2) Ijictv4n7spl_17.pdf (ripublication.com)

3) Gokce, E. (2020, January 10). “Predicting used car prices with machine learning techniques. “

4) Predicting Used Car Prices with Machine Learning Techniques | by Enes Gokce | Towards Data Science.

5) https://www.enjoyalgorithms.com/blog/car-resale-value-predictor-using-random-forest-regressor

6) Russell, S. (2015). Artificial Intelligence: A Modern Approach (3rd edition). PE

7) Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object