Targeted SUV Marketing using KNN

Dharamraj Bhatt
1947216, 4MCA
Christ (Deemed to be University),
Bangalore
dharamraj.bhatt@mca.christuniversity.in

Naishar Shah
1947258, 4MCA
Christ (Deemed to be University),
Bangalore
shah.vikram@mca.christuniversity.in

***Abstract*— The SUV segment is a significant growing market in this current era as it comes with practicality and space in town for day-to-day use. There is a company that wants to advertise their car based on the buyer's interest. The paper focuses on the K-Nearest Neighbor(KNN) Algorithm which is a type of predictive algorithm. The paper briefs about the KNN algorithm and further details. The dataset that has been referred to in this paper is SUV car data which contains multiple information about users. The paper illustrates the predictions of how many users are interested in buying a new SUV car so that company can send them ads. The KNN model has been trained on the training data and after training, the model is evaluated for test data.**

Keywords— KNN, K-Nearest Neighbor, SUV (Sport utility vehicle), prediction, machine learning, training, testing

I. Introduction

The aim of analytics is to make inferences from data using statistical and mathematical analysis. The analysis helps to make a prediction for some problem statements from the collected dataset. The solutions or other decisions can be provided with a data analytics tool like Online Analytical Processing (OLAP). Later it uses various tools and algorithms for better outcomes of data.

There are many technologies used in data analytics but predictive analytics is the one that uses machine learning algorithms and statistical analysis for future prediction. Here, we are going to use a KNN algorithm for SUV targeted marketing. Using this algorithm one can predict how many people are interested in buying that particular SUV model or not which is launched by the brand.

II. Problem Statement

There is a Car manufacturer company that has manufactured a new SUV car. The company wants to give the ads to the users who are interested in buying that SUV. So for this problem, we have a dataset that contains multiple user's information through the social network. The dataset contains lots of information but the Estimated Salary and Age we will consider for the independent variable and the Purchased variable is for the dependent variable.

III. KNN Algorithm

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on the Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a good suite category by using K-NN algorithm. The K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumptions 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.

Advantages: these are some of the advantages of the KNN algorithm.

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages: these are some of the advantages of the KNN algorithm.

- Always needs to determine the value of K which may be complex sometimes.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

IV. Dataset Description

This dataset consists of 5 attributes, this dataset is collected from social media users based on their interest and relevant questions like gender, age, etc.

This algorithm understands which user base should be targeted for SUV marketing of a brand with an accuracy rate of 93%.

Attributes information:

- User ID - Unique ID for each user
- Gender - Gender of the user.
- Age - Age of the user.
- Estimated Salary - Salary of the user.
- Purchased - 0 for users not interested and 1 for

interested people.

V. Model Construction

One of the main components of any data analytics problem is model construction. these are the steps we have followed to construct a model-

1. Data Preprocessing step
2. Fitting the K-NN algorithm to the Training set
3. Predicting the test result
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the train set result
6. Visualizing the test set result.

Data Preprocessing: As there was no null value in the dataset, preprocessing done to the dataset is-

- Extracting Independent and dependent Variable
- Splitting the dataset into training and test set.
- Feature Scaling

Age and estimated salary are considered as independent variables and purchased as the dependent variable.

.

Fitting the K-NN algorithm to the Training set: To fit the KNN algorithm to the training set we have imported KNeighborsClassifier. the result is shown below-

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                     weights='uniform')
```

Fig. 5.1: Fitting the training set

Predicting the test result: The predicting result is shown below-

```
[0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0
 0 0 1 0 0 0 0 1 0 0 1 0 1 1 0 0 1 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 0 1 0 0 1
 0 0 0 0 1 1 1 1 0 0 1 0 0 1 1 0 0 1 0 0 0 0 0 1 1 1]
```

Test accuracy of the result: Creating the confusion matrix by using two variables y_test and y_pred, the result is:

```
[[64  4]
 [ 3 29]]
```

Fig. 5.2 Confusion matrix

The accuracy of the test dataset result is **93%.**

Visualizing the train set result: To visualize the train set result we have used x_train and y_train and generated a listed color map graph and the result is shown below:

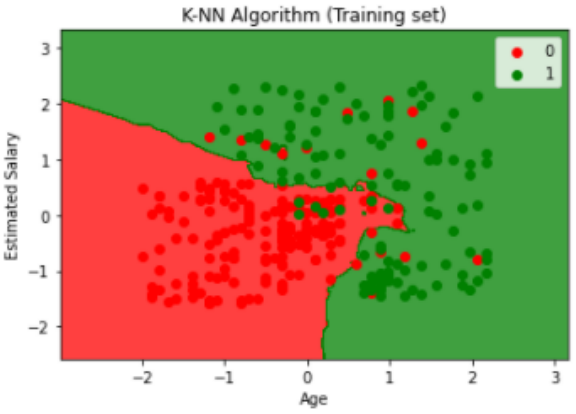

Fig. 5.3: Train set

It above graph can be understood by the below points:

- As we can see the graph is showing the red point and green points. The green points are for the Purchased(1) and Red Points for the not Purchased(0) variable.
- The graph is showing an irregular boundary instead of showing any straight line or any curve because it is a K-NN algorithm, i.e., finding the nearest neighbor.
- The graph has classified users in the correct categories as most of the users who didn't buy the SUV are in the red region and users who bought the SUV are in the green region.
- The graph is showing good results but still, there are some green points in the red region and red points in the green region. But this is no big issue as by doing this model is prevented from overfitting issues.

Visualizing the test set result: To visualize the train set result we have used x_test and y_test and generated a listed color map graph and the result is shown below:

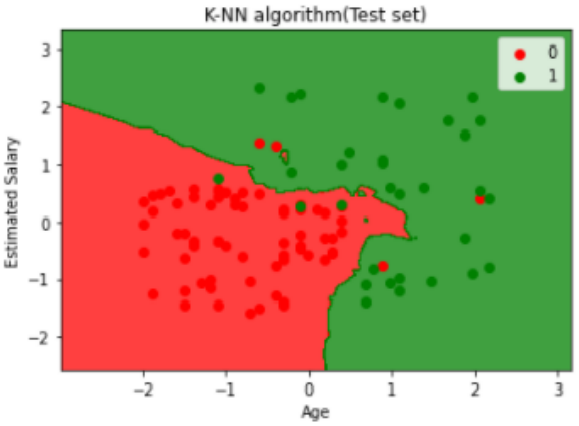

Fig. 5.4: Test set

The above graph is showing the output for the test data set. As we can see in the graph, the predicted output is good as

most of the red points are in the red region and most of the green points are in the green region.

VI. Inferences and Conclusion

After implementing the algorithm, the training dataset algorithm got an accuracy of 91% and the testing dataset algorithm got an accuracy of 93%. This is shown in the diagram below:

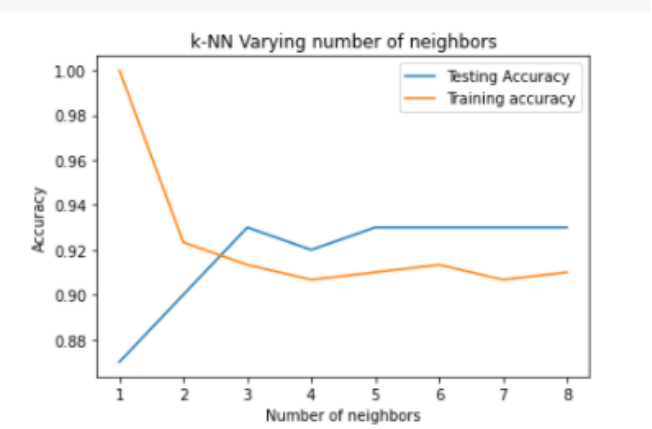

Fig. 6.1: Accuracy - train vs test

References

1. https://www.kaggle.com/iamaniket/suv-data
2. Guo, Gongde, et al. "KNN model-based approach in classification." *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, Berlin, Heidelberg, 2003.
3. Zhang, Shichao, et al. "Learning k for knn classification." ACM Transactions on Intelligent Systems and Technology (TIST) 8.3 (2017): 1-19.
4. Zhang, Shichao, et al. "Efficient kNN classification with different numbers of nearest neighbors." IEEE transactions on neural networks and learning systems 29.5 (2017): 1774-1785.
5. Models and Algorithms for Intelligent Data Analysis (Thomas A. Runkler)