A picture containing knife

Description automatically generated

Module 1 Assignment

**Classification Using Nearest Neighbors**

ALY 6020 – Predictive Analysis

Instructor – Justin Grosz

Name -Ashish Kothari NUID- 001360980

**INTRODUCTION**

K-Nearest Neighbor algorithm is a method for classifying objects based on the closest training examples in the feature space. KNN is a type of Instance – Based/ Lazy Learning where the function is only approximated locally, and all competition is delayed until classification. The KNN algorithm is fundamental and one of the simplest classification techniques when there is little or no prior knowledge about the distribution of the data. KNN is used mostly for classification but can be used for regression as well. We calculate the distance using Euclidean distance or Hamming Distance.

Applications:

1. Weather forecasting
2. Text mining
3. Stock market forecasting or uncovering market trends

Advantages:

1. It is robust in noisy training data
2. It is effective if the training data is large
3. There is no training phase

Disadvantages:

1. We have to pass the value of K and it is difficult to apply in high dimensional data
2. We can get false output if distance is very close or almost on the same point
3. It is difficult to understand which type of distance metric to use

In this project we are going to take NFL match situation under consideration and apply KNN method. We will use Football.NN dataset that consists of 15 different variables and we are going to predict yards\_gained for NFL Running Plays.

Steps:

1. We are going to limit our model to only 5 variables using exploratory analysis.
2. We will build model for values of k= 1,2,5,10
3. Then using feature importance, we are going to compare the accuracy of two models
4. We will determine which variables are most significant to predict the success of running plays

Step 1:

First, we import the dataset and perform exploratory analysis using heatmap.

Graphical user interface, text, application

Description automatically generated

Chart, timeline, waterfall chart

Description automatically generated

In this heatmap we have plotted all the variables and then we see which variables have the maximum importance with the predict data (yards\_gained). We notice that the variables yardline\_100, ydstogo, shotgun, qb\_dropback and Run gap End. So, we have chosen our 5 best variables based on the exploratory analysis. Using these 5 variables we can do the KNN method and see what accuracy is of using variables for k= 1, 2, 5, 10.

Step 2:

We split the data into training data and test data. We split the dataset by 80% of training data and 20% of test data. Using the method of KNeigborsClassifier we create the KNN score.

For k= 1, 2, 5, 10 we get the accuracy of 70%, 64%, 60% and 54% respectively.

Graphical user interface, text, application

Description automatically generated

Step3:

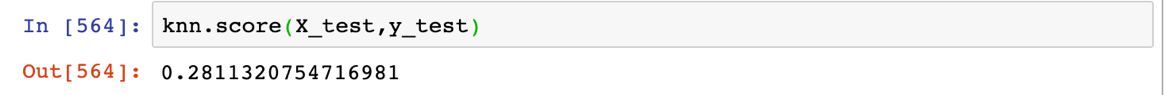
Now we select all the variables for feature importance and create a bar/ Histogram to check which are the best 5 variables we should use. We plot the model using RandomForest and get a bar plot.

Chart, histogram

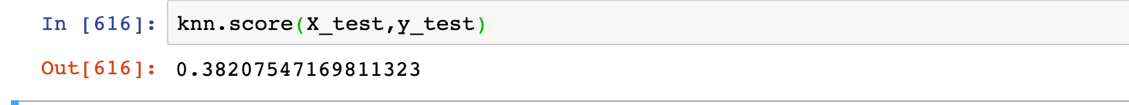
Description automatically generated

From this bar plot we can see the most important five variables are yardline\_100, game\_seconds-remaining, drive, ydstogo and shotgun. We notice that three variables yardline\_100, ydstogo and shotgun are repeated.

We create KNN model using these 5 variables and the accuracy for k= 1, 2, 5, 10 are 38%, 34%, 31% and 28% respectively.









Step 4:

We compare the accuracies of the KNN using exploratory analysis and feature importance, we notice that the accuracy with exploratory analysis is very high as compared to accuracy with feature importance. We can say that the variables from exploratory analysis yardline\_100 ydstogo, shotgun, qb\_dropback and Run gap End are important and should be considered to plan the play accordingly.

Step 5: Conclusion

After comparing and seeing the accuracies we can assume that coach should focus on five things yardline\_100, ydstogo, shotgun, qb\_dropback and Run gap End. Seeing these conditions the coach can plan their game play and focus more on shotgun type of play, qb\_dropback and look for gaps in the end. This will give them a higher chance to gain maximum number of yards or we can use shotgun game play and qb\_dropback game play. These two game plays can increase chances of covering more yards. Coach should focus on these for the next upcoming season of NFL.