Project 3: Classification Models Report

CSC 177-01: Data Warehousing and Data Mining

Professor: Jagan Chidella

Group: Jason Phillips, Mohammad Ameri, Ryon Faroughi, Youser Alalusi, Yusran Sadman

Introduction

The name of the dataset we're using is "Churn_Modelling", it is the percentage of subscribers to a service who discontinue their subscription to the service within a given time period. For a company to grow the amount of their customers, its growth rate, as measured by the number of new customers, must exceed its churn rate. Based on the person's credit score, region, gender, age, tenure, balance, and salary we can figure out whether or not the person will EXIT(1) or NOT(0). Also we perform preprocessing techniques on the rest of the columns and split the data up into the standard 80/20 for the training data and testing data.

Shuffling data serves the purpose of reducing variance. Normalizing the data is to reduce data redundancy. Dropping useless columns such as RowNumber, CustomerId, Surname. Also, One Hot encoding is necessary because encoding features with a range of numbers isn't really helpful.

Naive Bayes

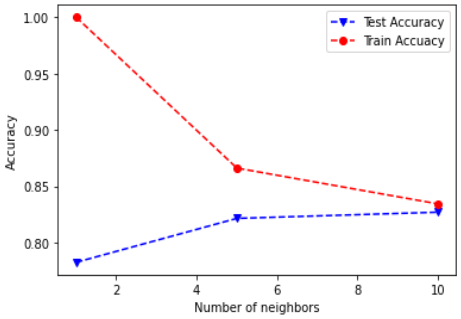
Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Gaussian Naive Bayes is implemented for classification.

```
#---Naive Bayes---

clf_NB = GaussianNB()
clf_NB.fit(trainX,trainY)
NB_pred = clf_NB.predict(testX)
print('Accuracy on test data is %.2f' % (accuracy_score(testY, NB_pred))

Accuracy on test data is 0.83
```

Naive Bayes Accuracy is 83%

K-Nearest Neighbor

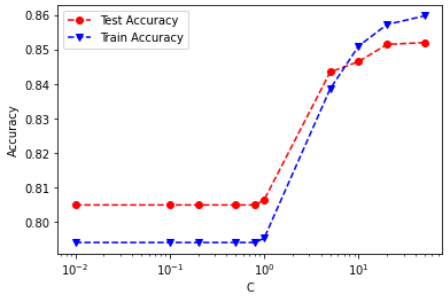
K-Nearest Neighbors, or KNN for short, is one of the simplest machine learning algorithms and is used in a wide array of institutions. KNN is a non-parametric, lazy learning algorithm. When we say a technique is non-parametric, it means that it does not make any assumptions about the underlying data. In other words, it makes its selection based off of the proximity to other data points regardless of what features the numerical values represent. Being a lazy learning algorithm implies that there is little to no training phase. Therefore, we can immediately classify new data points as they present themselves.



K-Nearest Neighbor Accuracy is 83%

Support Vector Machines

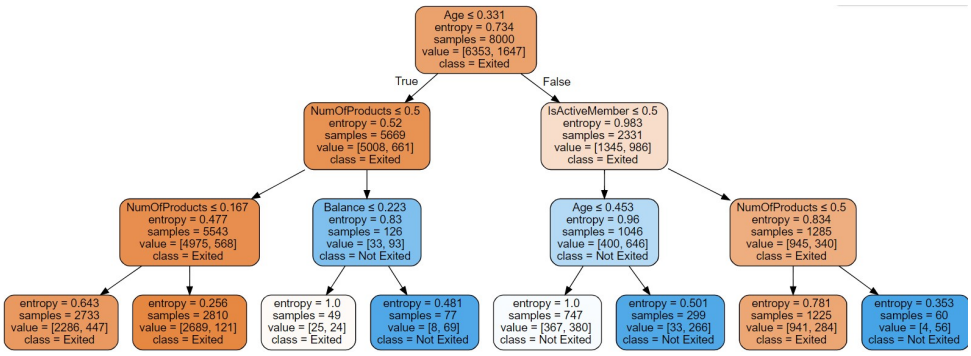
Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes. Here you'll see a non linear decision boundary, which shows the larger the C value the more time it needs to run.



Accuracy is 85%

Decision Trees

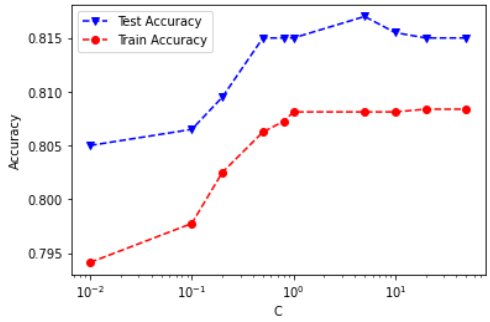
Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.



Accuracy is 82%

Logistic Regression

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.



Accuracy is 81%