**Summary:**

Our goal is to predict whether a customer will make a term deposit through exploratory data analysis and machine learning on a 45,211 observation dataset that includes demographic data and prior marketing campaign responses.

K Nearest Neighbor (KNN), RandomForest (RF), and Support Vector Machines Classifier(SVC), classification models were run on the dataset. The data was spilt into 80/20 training testing sets. StandardScaler(), PCA(n\_components=7), and 5 fold cross validation were applied to the data prior to running classification models. KNN with k\_neighbors = 18 was the optimal model. It had the highest cross validation score of 88.92% and a predictive accuracy of 86% on the test data. In addition, its accuracy was consistent among training and testing data sets.

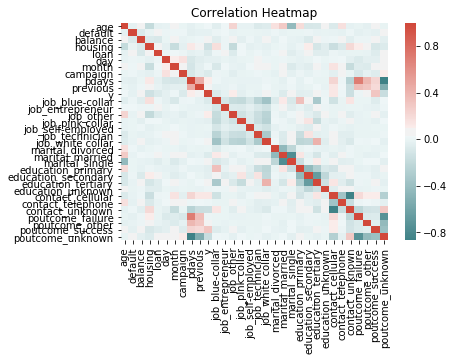
**Data Source**: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

**Data exploration**:

**Data cleansing / pre-processing**:

1. Reduce the number of categories for jobs to simply model ( e.g. adding management jobs to blue collar category)
2. Map categorical data to integer value for the following attributes (month, default, housing loan, personal loan, y(deposit) )
   1. Yes no binary mapping (1 = yes, 0 = no)
3. Dummifying remaining categorical attributes. Dummifying changes categorical attributes into several columns with binary values for each category. For example, the job attribute has several categories (white-collar, blue-collar, service) and the “get dummies” function creates a columns for each category with a 1 and 0 as a result( 1 = True, 0 = False).
4. Create test train split for machine learning.

**Correlation/Covariance analysis**:

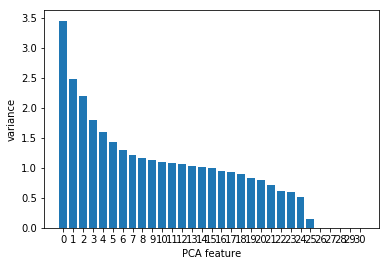


|  |  |
| --- | --- |
| Attribute | Deposit Correlation |
| poutcome\_success | 30.68% |
| contact\_cellular | 13.59% |
| pdays | 10.36% |
| job\_other | 10.29% |
| previous | 9.32% |
| education\_tertiary | 6.64% |
| job\_blue-collar | -7.21% |
| campaign | -7.32% |
| housing | -13.92% |
| contact\_unknown | -15.09% |
| poutcome\_unknown | -16.69% |

The attributes with the strongest correlations were (poutcome = true) and (contact = ‘cellular” ) with 30.7% and 13.6% correlation, respectively. These correlations are logical as most people use cell phones as their main contact method and if prior marketing campaigns were successful, the next marketing campaign is more likely to be successful. Variables were checked for high covariance to prevent overfit, but there were not any.

**Exploratory Data Analysis**:

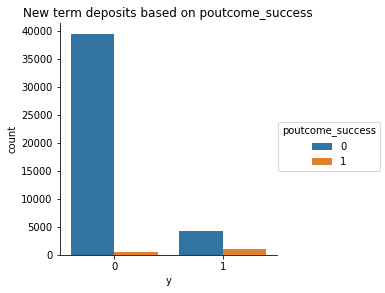
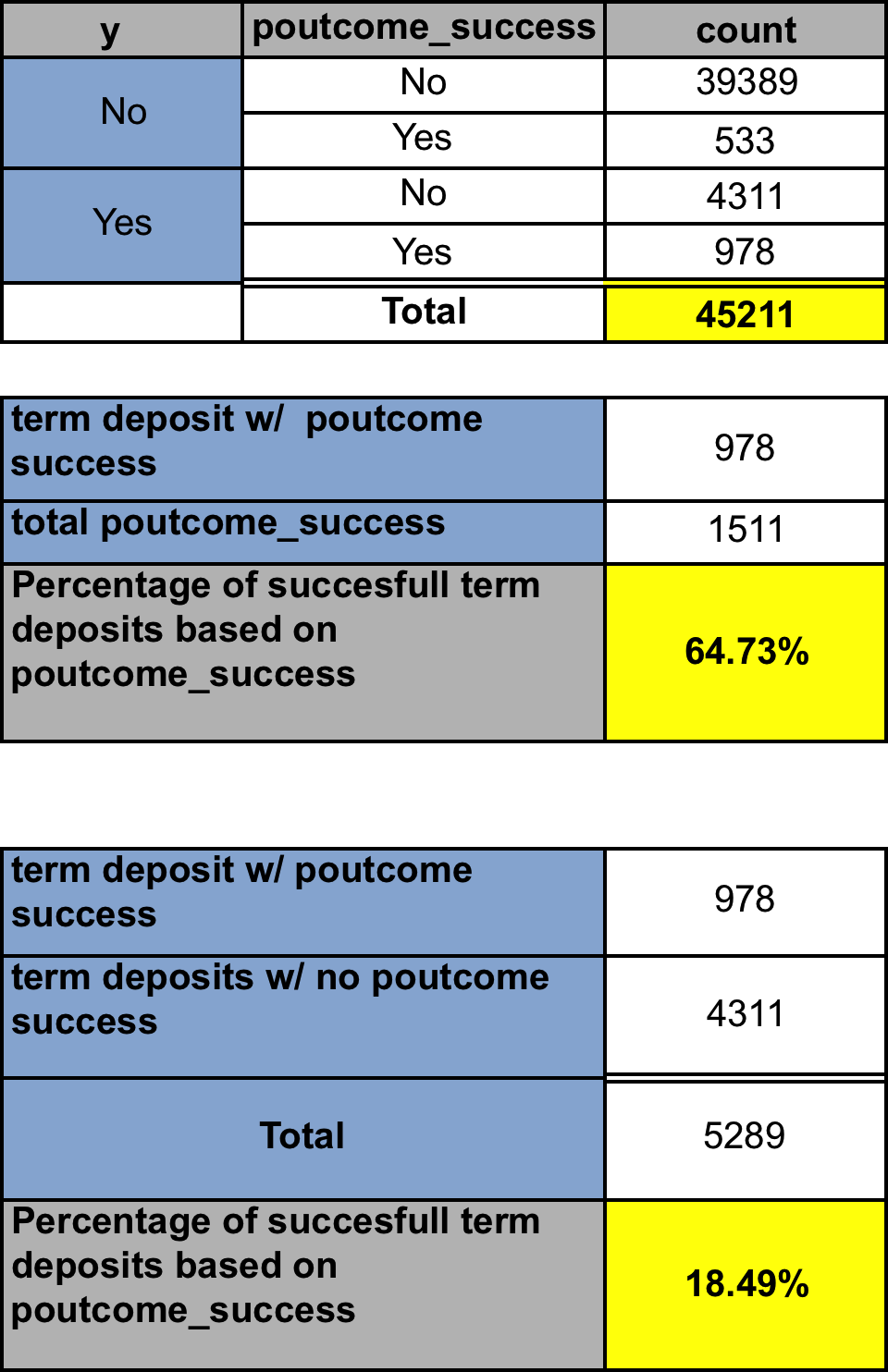
**Principal component analysis**



StandardScaler() needs to run on the data frame features prior to PCA.

PCA identifies important features, which are features that have high variances. There are 7 features that have large variances. As a result, PCA was set to PCA(n\_components = 7) for our classification models.

**Poutcome\_success**

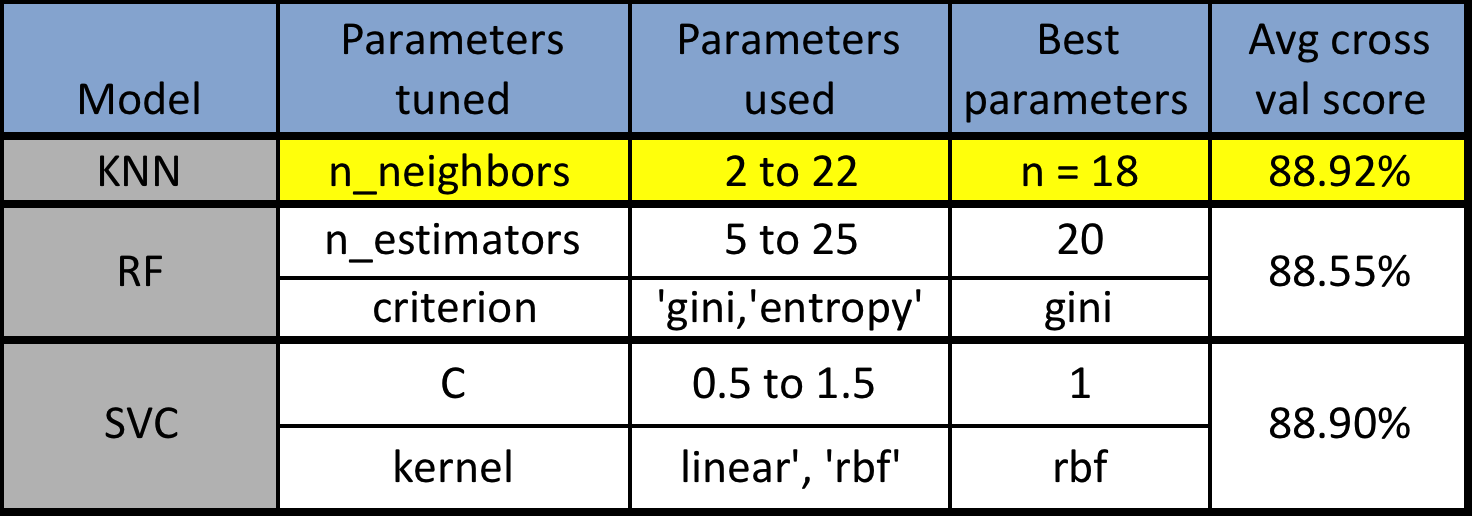


**Predictive Classification Model Selection and Tuning**

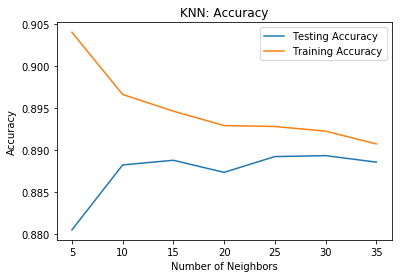
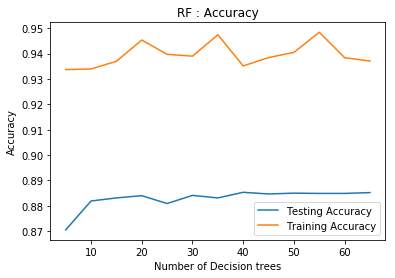
Poutcome\_successful was an effective predictor of term deposits with a success rate of 64.7 %.

18.5 % of term deposits had outcome\_success.

**Training data tuning accuracy metrics**

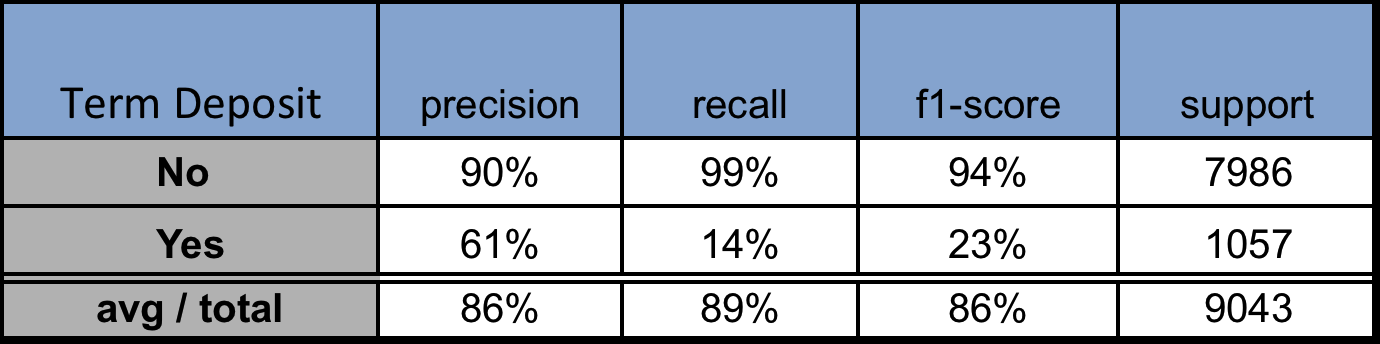


**Training vs testing accuracy**:



KNN is the optimal model because it has the highest training data cross validation accuracy metrics and its training and testing accuracy scores are the most consistent. RF and SVC have bigger gap between training and testing accuracies.

**KNN testing prediction metrics**



The KNN prediction model had 86% accuracy.