DEPARTMENT OF COMPUTER ENGINEERING,

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

SAVITRIBAI PHULE PUNE UNIVERSITY

Academic Year: 2020-21

Semester: I

A Mini-Project Report

On

BANK CUSTOMERS EXITNG PREDICTIVE MODEL

By

VARUN GADDE - BECOB203
VEDANT UPGANLAWAR - BECOB270

Guided by
Prof. Priya Surana

Table of Contents

Report on Mini Project

Title:

A bank is investigating a very high rate of customers leaving the bank. Investigate and predict which of the customers are more likely to leave the bank soon.

Introduction:

This project focuses on applying suitable data preprocessing steps such as handling of null values, data reduction, discretization of the 'Churn_Modelling.csv' dataset having 10,000 records of customers of a bank.

Aim is to Investigate and predict which of the customers are more likely to leave the bank soon by using different techniques - Logistic Regression, Support Vector Machine, Naive Bayes Classifier, KNN and Random Forest Classifier and further analyze the confusion matrix and compare these models along with applying cross validation while preparing the training and testing datasets.

Requirements:

The project requires the following imports:

1. Numpy - for Linear Algebra
2. Pandas - for Data Preprocessing and CSV I/O
3. Matplotlib - Data Visualization
4. sklearn.model selection - for Modelling
5. sklearn.linear_model - for Logistic Regression Classifier
6. sklearn.svm - for Support Vector Classifier
7. sklearn.naive_bayes - for Gaussian Naive Bayes Classifier
8. sklearn.neighbors - for KNeighbors Classifier
9. sklearn.ensemble - for Random Forest Classifier
10. sklearn.metrics - for Accuracy Score, Confusion Matrix and Classification Report

Theory:

CLASSIFICATION

Classification is a data mining technique that assigns categories to a collection of data in order to aid in more accurate prediction. Classification is one of several methods intended to make the analysis of very large datasets effective. The goal is to create a set of classification rules that will answer a question, make a decision, or predict behavior. To start, a set of training data is developed that contains a certain set of attributes as well as the likely outcome.

The job of the classification algorithm is to discover how that set of attributes reaches its conclusion.

It is a two-step process such as:

Learning Step (Training Phase): Construction of Classification Model Different Algorithms are used to build a classifier by making the model learn using the training set available. Model has to be trained for prediction of accurate results.
Classification Step: Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.

Different Classifiers used in Machine Learning:

1. Decision Trees
2. Naïve Bayes Classifiers
3. Neural Networks
4. K-Nearest Neighbor
5. Support Vector Machines
6. Linear Regression
7. Logistic Regression
8. Random Forest Classifier

Algorithms Used in this project:

1) Random Forest classifier.

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well.

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions.

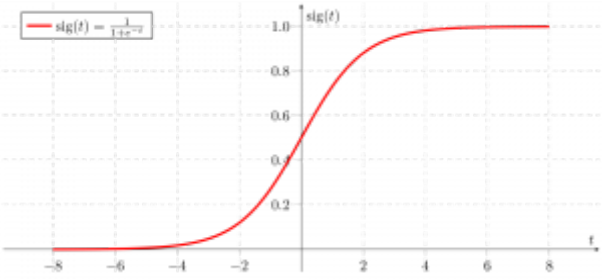
The reason for this wonderful effect is that the trees protect each other from their individual errors.

So the prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.
2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

2) Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes.



Logistic regression models the data using the sigmoid function.

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.

3) Support Vector Machine

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where the number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

4) 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.

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

5) KNN

K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.

KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. In other words, the model structure determined from the dataset. This will be very helpful in practice where most of the real world datasets do not follow mathematical theoretical assumptions. Lazy algorithm means it does not need any training data points for model generation. All training data used in the testing phase. This makes training faster and the testing phase slower and costlier. Costly testing phase means time and memory. In the worst case, KNN needs more time to scan all data points and scanning all data points will require more memory for storing training data.

ATTRIBUTES ANALYSIS

On analysis of the attributes of dataset, attributes are categorized as follows:

1. **Categorical Variables**
   - Geography
   - Gender
   - HasCrCard
   - IsActiveMember
   - Exit (determines whether the customer left the bank or not)
   - Num Of Products
2. **Numerical Variables**
   - CreditScore
   - Age
   - Tenure
   - Balance
   - EstimatedSalary

PREPROCESSING

The following data preprocessing techniques are applied in the project.

1. Missing Values
2. Dropping of Unnecessary Attributes
3. Discretization

MODELLING

Modelling includes creation of train, test data by splitting the dataset.
Splitting the X_train and y_train as X_Train,y_Train,X_Test,y_test.
The split ratio varies according to the discretion of the user.
Here we have used the splitting ratio as 0.33.

Comparison of all algorithms on Churn_Modelling.csv Dataset:

Classifier	Accuracy (Test Data)	K-Fold Cross Validation Accuracy	Confusion Matrix	Classification Report
Random Forest (RF)	86.51	85.31	[[2554 103] [342 301]]	precision recall f1-score support 0 0.88 0.96 0.92 2657 1 0.75 0.47 0.57 643
Logistic Regression (LR)	85.0	83.61	[[2552 105] [399 244]]	precision recall f1-score support 0 0.86 0.96 0.91 2657 1 0.70 0.38 0.49 643
Support Vector (SVM)	84.66	83.46	[[2602 55] [451 192]]	precision recall f1-score support 0 0.85 0.98 0.91 2657 1 0.78 0.30 0.43 643
Naive Bayes (NB)	82.78	81.56	[[2638 19] [549 94]]	precision recall f1-score support 0 0.83 0.99 0.90 2657 1 0.83 0.15 0.25 643
K Nearest Neighbour (KNN)	82.0	80.20	[[2559 98] [502 141]]	precision recall f1-score support 0 0.84 0.96 0.90 2657 1 0.59 0.22 0.32 643

K- FOLD CROSS VALIDATION

In this case of cross validation, we will divide the complete dataset into k parts.Each section will act as the test set and rest of the k-1 sections will be used for training.Hence we get k different accuracy scores.For this dataset, we have **k=10**.Finally from these k scores, we obtain a mean accuracy which helps us compare the classifiers with better precision.

PROJECT LINK:

https://colab.research.google.com/drive/12v1-kAoj6jF4YvWxT4Y65GPhnGlFRRT6