**<Bank Marketing Project Overview>**

**201533661이승수, 201533645배성재**

**<Contents>**

1. Objective Setting: set Marketing Strategy for increasing term-deposit subscribers.

Sub-Objective: Predict ‘y’ variable(if client subscribed a term deposit:yes/no)

1. Data Curation: bank.csv
2. Data Inspection
3. Data Preprocessing
4. Data Analysis
   1. Decision Tree, Logistic Regression, SVM
5. Evaluation

Evaluation methods: k-fold cross-validation, compare result with Wisephropet

1. Deployment
2. **Objective Setting**

Our object is to predict ‘y’ feature which means whether client subscribed a term deposit, and through predicting, we want to look for which features are affects to result.

Our final goal is to find out affectable key features usable for term deposit service marketing.

1. **Data Curation**

UCI: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

1. **Data Inspection**

‘bank.csv’ data is first 10% dataset of the ‘bank-full.csv’ data for training.

Firstly, we split data ‘bank.csv’ by ‘;’ notation and make into ‘bank-preprocessed.csv’.

We looking for dataset ‘bank-preprocessed.csv’. features of the data is as below.

Row: 4521 column: 17

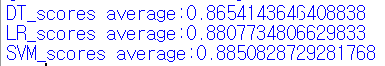
|  |  |  |
| --- | --- | --- |
| Feature | definition | Data type |
| Age |  | numeric |
| Job | type of job | categorical |
| Marital | marital status | categorical |
| Education |  | categorical |
| Default | has credit in default | binary |
| Balance | average yearly balance, in euros | numeric |
| Housing | Does has housing loan | binary |
| Loan | Does has personal loan | binary |
| Contact | contact communication type | categorical |
| Day | last contact day of the month | binary |
| Month | last contact month of year | categorical |
| Duration | last contact duration, in seconds | numeric |
| Campaign | number of contacts performed during this campaign and for this client | numeric |
| Pdays | number of days that passed by after the client was last contacted from a previous campaign | numeric |
| Previous | number of contacts performed before this campaign and for this client | numeric |
| poutcome | Does client subscribed a term deposit | binary |

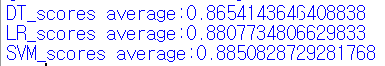
After looking data, we decided to change some data types like ‘month’(categorical to numeric), and decided there’s no dirty data to be cleansed.

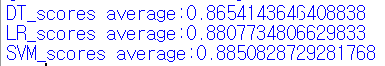
1. **Data Rreprocessing**

After data inspection, we make machine learning model to predict target ‘y’ feature and then want to looking for which features matters by plotting result after manipulating features one by one.

We choosed Decision Tree Classification, Linear Regression, and SVM algorithms for predicting ‘y’ feature. Accuracy is as below.

**Decision Tree:** 

**Linear Regression:** 

**SVM:** 

Before feature engineering, SVM shows best accuracy about 88.5%.

Then we plot prediction results by X labels as rest of each 16 features and Y labels as frequency of original ‘y’ value. If ‘y’ is ‘yes’, plot with blue lines and plot with red if ‘y’ is ‘no’.

For plotting, we have to devide features into categorical/numeric features and plot them because at plotting categorical values, X labels have to be in ordered. So we use subcategorybar() function in plotting categorical values by function plotColdNY\_categorical().

Then, we look for each plots’ distribution and NYratio which means Yes/No ratio at ‘y’ for each features. We pick affectable features by predictors those have high NYratio.

**NYratio(Yes/No Ratio) =**

Plotting results of each features are as below.

**[Plots for each features for feature selection]**

|  |  |
| --- | --- |
| 'age' | Frequency shows highest at ‘age’=35 and distribution of data is focused on ‘age’ is about 30 to 60. |
| 'job', | ‘job’=management shows highest frequency.  But **‘age’=student, retired, unknown** shows highest nyRatio regardless of frequency. So it can be considered as affectable factors for prediction. |
| 'marital', |  |
| 'education', |  |
| 'default', |  |
| 'balance', | ‘balance’ is continuous data, but goes abnormally high at ‘balance’=80, to 357. And value with ‘balance’>4000 looks like outlier. |
| 'housing', |  |
| 'loan', |  |
| 'contact', |  |
| 'day', |  |
| 'month', | ‘month’=6, which means June have largest data but it doesn’t shows that much ‘yes’ frequency, so **‘month’=June have negative affect for ‘y’ prediction.** |
| 'duration', | Values of ‘duration’s features with over 700 shows higher nyRatio than average. So I **binned ‘duration’ values devided by 10 and 100.**    This is plot of ‘duration’ feature, which’s value is binned by 1/10.    This is plot of ‘duration’ feature, which’s value is binned by 1/100.  We will use **1/100 binned ‘duration’ feature** as a result. |
| 'campaign', |  |
| 'pdays', | ‘pdays’=-1 means client was not previously contacted      Distribution of ‘pdays’ data is shown as above. |
| 'previous', | Distribution of ‘previous’ feature shows most of them at ‘previous’=0, but correlation doesn’t show int nyRatio. |

|  |  |
| --- | --- |
| 'poutcome', | ‘poutcome’=success don’t show that much frequency, but nyRatio shows abnormally high, which is about 6 times of another alternatives. |

As a result, we can conclude affectable features as ‘job’, ’poutcome’,’ ‘month’, and ‘duration’.

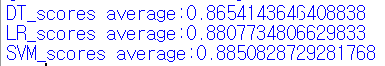
Positive/negative relationships for each features are as below.

|  |  |
| --- | --- |
| **Feature** | **Correlation with Yes/No\_Ratio** |
| **Job** | **Student, retired, unknown** |
| **Poutcome** | **success** |
| **Month** | **6(June)** |
| **Duration** | **700+,** |

1. **Data Analysis**

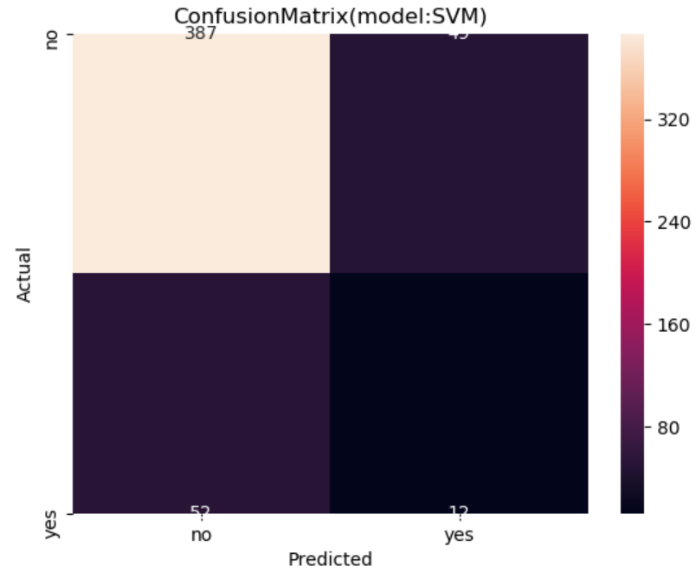
Average accuracy of classification goes up from about 83% to 87% after the feature engineering.

-Average Accuracy(cross-fold-validation accuracy) before feature engineering: 83%



-Average Accuracy(cross-fold-validation accuracy) after feature engineering: 87%



1. **Evaluation**

Using confusion matrix, we can get precision and recall as below.

|  |  |
| --- | --- |
| FP=52 | TP=12 |
| TN=387 | FN=45 |

Then,

Precision=12/64= 0.187.

Recall=12/57=0.210 .

And using Wisephropet, we can find that top 5 affective features matches 3 of our selected features, so we can conclude that we select 4 important features and three of them are most affective features. It means prediction

|  |  |
| --- | --- |
|  |  |
| **<my selected most important features>** | **<Wiseprophet’s selected most important features>** |