**Project diary**

|  |  |
| --- | --- |
| **Week** | **1** |
| **Communication** | * MS Team meeting Sat 28th March 2020 * MS Team chat |
| **Any analysis or other difficulties that arose and how they were overcome** | * Conducted exploratory analysis of data and looked at the integrity of the bank-full-additional dataset (missing values, errors etc. in SAS. |
| **Progress made** | * Discussed the dataset and various data resources * Decided on a final deadline for coding and write-up * Set agenda for week 1 |
| **Member contributions to progress** | * Amol - Explore and dataset understanding, Created Group in MS Teams, conducted project meeting, Created GitHub repository for project * Anurag - Explore and dataset understanding in SAS. * Bindu – Analyzed the data in R (Checked for missing values) * Alex – Figure out how to create a basic logistic model in SAS * Jyothi - Explore and dataset understanding * Luke - Screened data and checked dataset integrity using SAS (Outlier/extreme value detection, missing values detection) * Nishant - Explore and dataset understanding |

|  |  |
| --- | --- |
| **Week** | **2** |
| **Communication** | * MS Team meeting Sat 4th April 2020 * MS Team chat |
| **Any analysis or other difficulties that arose and how they were overcome** | * Analyzed dataset and found out that it includes missing values and non-numeric categorical predictors. * Not all predictors are important/influential. * After EDA, we will be able to decide more in details. |
| **Progress made** | * Discussed on the dataset, and difficulties faced. Agreed on the next week goal, to do the pre-processing step of the project. * Decided on to do the pre-processing i.e., data cleaning and manipulation in SAS, R and python separately in the groups of 2 or 3. |
| **Member contributions to progress** | * Amol - Dataset exploration, proposed the steps to follow to complete the project, [click here](https://github.com/NanawareAmol/Project_Portuguese_Banking/blob/master/Steps.txt) for more details. * Anurag - Dataset exploration, detection of Outliers, generating Graphs for Summary Statistics. * Bindu - Started data tidying in R (Converted categorial variables to numeric and checked for correlation) * Alex - Further scrutinized logistic model with backwards elimination. * Jyothi - Dataset exploration and data preparation * Luke - Began data cleaning and pre-processing in SAS (Encoded categorical variables as numeric, outlier/unknown value deletion, correlation checking) * Nishant - understand the different columns of dataset by creating visuals to help the application of logistic regression in the dataset. |

|  |  |
| --- | --- |
| **Week** | **3** |
| **Communication** | * MS Team meeting, Sat 11th April 2020 * MS Team chat * SAS group meeting on 13/04/2020 * SAS group meeting on 17/04/2020 * Python group meeting on 15/04/2020 |
| **Any analysis or other difficulties that arose and how they were overcome** | * Analyzed different techniques on data cleaning * Analyzed and decided on creating dummy variables for each value in categorical variables instead of assigning weights to the values. * Got different number of columns after cleaning data for each software. |
| **Progress made** | * Discussed the dataset, and difficulties faced while cleaning the data. * Set agenda for week 3 - complete pre-processing steps and fitting the model. |
| **Member contributions to progress** | * Amol - Cleaned the data in two ways, one with dropping all missing values and another one with imputing missing values. Fitted logistic regression on the full dataset in python. * Nishant - Pre-processing and model fitting in Python. * Anurag - Data pre-processing, checked for Correlation and deleted columns, accuracy comparison on Training & Testing datasets. * Luke - Explored methods of further cleaning data (SMOTE to account for class imbalance, multicollinearity checking. Partitioned dataset while maintaining response class ratio.) * Alex - Work on model scoring and comparison with ROC curves. * Bindu - Fitted logistic regression on the full dataset in R. R fails elimination method (wants dataset with 0 NA’s). Imputed few missing columns and removed few. * Jyothi - Data preparation and model fitting in R |

|  |  |
| --- | --- |
| **Week** | **4** |
| **Communication** | * MS Team meeting, Sat 18th April 2020 * MS Team chat |
| **Any analysis or other difficulties that arose and how they were overcome** | * Compared the model accuracy using cross validation between SAS, python and R. * Got different accuracy in all software, since the data processing is different in the respective libraries in each software. |
| **Progress made** | * Discussed the dataset, and difficulties faced fitting the models based on validation. * Set agenda for week 4 – comparison of model-fitting across all three software and came up with pointers for final report. |
| **Member contributions to progress** | * Amol - Performed model fitting in python on reduced logistic model(Recursive feature elimination). Created performance metric using k fold cross validation. * Nishant - created visualization with different combination columns to give a clear understanding . * Anurag - Fitted Logistic Regression Model with different methods like Backward, Forward and Stepwise elimination methods to find the most optimal predictors. * Luke - Added to SAS code and group report. Transferred project diary from Amol’s GitHub into correct template and added to it. * Alex - Compiled SAS code and worked on group report * Bindu - Performed model fitting in R. (Plotted ROC curve, calculated the accuracy, confusion matrix and MSE of the model) * Jyothi - Fitted different logistic regression models in R and analyzed the result based on accuracy, confusion matrix and AUC scores in R |

|  |  |
| --- | --- |
| **Week** | **5** |
| **Communication** | * MS Team meeting, Sat 25th April 2020 * SAS group meeting, Sat 25th April 2020 * MS Team chat * Emailed Rafael for clarification 28th April 2020 |
| **Any analysis or other difficulties that arose and how they were overcome** | * Compared the model accuracy using cross validation between SAS, python and R. |
| **Progress made** | * Discussed the difficulties faced fitting the models based on validation and final report. * Set agenda for week 5 – Complete and submit project. |
| **Member contributions to progress** | * Amol - Enhanced the model with removing high correlation and created few plots with feature importance. * Nishant - Checked on the reduced model with high correlation and created few plots for observation. * Anurag - Had discussion on the Final Report, further checked the Dataset for certain anomalies. * Luke - Created confusion matrix in SAS. Reformatted and added to group report. * Alex - Edited and compiled the final report * Bindu - Performed model fitting for python cleaned dataset. Compared the results of different fits. * Jyothi - Fitted new model with categorical variables encoded to dummy and compared the results with the existing models in R |