**Individual Reflective Report – Priya Arvind**

For this assignment, I and my colleague Ivan decided to adopt the Cross-Industry Standard Project for Data Mining (CRISP-DM). We agreed on a division of duties according to the different phases of the CRISP process. I was in charge of the Data understanding (EDA/PCA), Model Building phase for building the predictive model for the target feature-defaulter and contribute Conclusion/Next Steps with the team.

My contribution to the project is detailed as follows:

* Understand the dataset with the help of different visualization techniques. This included the following tasks:
* Data visualization techniques like pair-plot, histograms, heat map, Missingno, Correction Mapping and libraries were used.
* Applied PCA on the pre-processed data to understand the data distribution and correlation between the data points.
* Evaluate the data on Over-Sampling techniques and applied the Over-Sampling SMOTE technique on scaled Data points for Classification Model, since the target variables were highly imbalanced.
* Models building. This phase included the following tasks:
* Since we have 2 different predictions, I tried to apply different Models for the target value-defaulter and considered the best models based on the evaluation done by the team.
* Classification: I applied 2 different Classification models- Logistic Regression and Random Forest Model on the dataset. Then applied Gird SV to find the best hyper Parameter to apply the best score parameters to each classification Model.
* In the group report, I prepared the Introduction- Objectives, Data understanding-EDA, model building for defaulters (Logistic Regression and Random Forest) and equal team contribution in the conclusion section.
* In the slide deck, I prepared the Data understanding -EDA, Model building and equal team contribution in the conclusion section and Introduction section.
* In this project, I have developed a deeper understanding of EDA and how simple visualization techniques can be used to understand the data and learned how to use GitHub repository.
* Overall we did good teamwork and try to explore and help each other to develop this project better.

**Individual Reflective Report – Ivan Saravanja**

Working on this assignment helped me to get the better understanding of how machine learning methods work and to realise the importance of the data preparation process.

My role in this assignment involved finding appropriate dataset to use, finding, and loading required libraries, importing, and merging datasets, pre-processing the data (feature selection, removing NaN values, removing duplicates, converting negative to absolute values, and replacing categorical variables with numerical values), building two machine learning models and creating parts of the report and PowerPoint presentation for the CA1and then simplifying and improving workbook in CA2 by applying newly acquired knowledge. Parts I did in CA2 were:

* Making the workbook less heavy on code – I tried to get clean, well explained workbook.  
  By doing that, I removed some parts from CA1, mostly those that refer to explaining datasets 1 and 2, as that wasn’t that important, since I shifted focus on merged dataset and all the codes later on were referring to merged dataset only.
* Improving visualisations – visualisations we had in CA1 were basic ones, few of them were unreadable as well due to size of dataset and number of columns. In this CA2, all the plots are helping to understand dataset better and bringing more insights. I used new learned skills such as density plot or interactive plot
* Data preparing – data preparation was also improved in CA2, which eventually led to quite improved scores for machine learning models. I did Standard Scaling and PCA. When executing PCA, I used cumulative variance ratio to determine the number of components to retain in my PCA to get at least 95% of variability in our data. In CA1 we were using 20 components, but in CA2, by applying cumulative variance ratio, we were using only 4 components.
* Improving Machine Leaning Models:
  + Decision Tree Model: While improving this model, I used RFE method to rank the importance of features in my dataset and, at the end, I was using only features that are within top 3 in the importance. Also, I used Gini to split the data in a way that minimizes impurity at each node. Other change compared to CA1 is that I applied 80/20 split, on top of 70/30 one that we used in CA1.   
    By doing all of these, scores improved just a bit:  
    - CA1 – 87.60%  
      CA2 – 87.69%
  + Logistic Regression: As mentioned above, what I did first to improve preparation of the dataset by performing Standard Scaling and PCA. What I also did was to use pipeline, for multiple reasons – helps in organising and streamlining ML workflow but it’s also useful when working in collaborative setting, like we were doing it now, as whole ML process can be reproduced consistently.   
    Scores of this model improved significantly:
    - CA1 – 60.78% even after using GridSearchCV  
      CA2 – 92.06%
  + Random Forest: This model reached the best accuracy among models I was working on. With above mentioned pre-steps, this was the change:
    - CA1 – 76%  
      CA2 – 92.02%
* Adjusting assignment report, personal report, and presentation.

To sum it up, this assignment gave me valuable insights into analysing loan applications, preparing data, and building predictive models. I learned the importance of focusing on data quality and implementing models effectively. This experience has deepened my understanding of the challenges faced in real-world data analysis.