

Group Number	20
Student ID	25269933
Name	Kesavarapu Vivek Reddy
Overall Percentage Contribution to the project	33.33%

Date: 22-10-25

Today I started to go through the project brief and understand the problem statement. I got the credit card default dataset from the UCI Repository and went over the available documentation to learn about the variables and structure.

Date: 24-10-25

Our initial team meeting was conducted to allocate duties for the project. We discussed our workflow strategy and agreed upon a gradual pipeline from EDA to modelling. I wrote down the tasks that I would be working on later, mainly feature-target analysis and model evaluation. After that I looked into the dataset size and the columns.

Date: 26-10-2025

Today, along with team we meet our professor Annette via Teams. We displayed the dataset to her and explained its organization and the key features. For further steps, she gave us feedback and some recommendations.

Date: 29-10-2025

I revisited to understand the dataset to understand missing value patterns, numeric ranges, and distribution issues. This was a preliminary scan only, but it nonetheless gave me a more precise idea of the complexity of the data. I made notes for our upcoming team meeting regarding possible difficulties.

Date: 31-10-2025

Today the updates of descriptive statistics and categorical exploration came from our team members Ben and Hima, and then I decided to dedicate a few hours to analyzing the behavior of numerical features so that I could get ready for my future work. The distribution of the EDA tasks was also finalized, aligned with our project timeline, and my notes were updated as well. I created a colab/Jupyter notebook structure and went through the different ways of visualizing correlations, etc.

Date: 04-11-2025

Joined another team discussion to review Ben's progress on histograms and bar plots. We discussed the early patterns in the dataset and identified that certain payment delay features might be strong predictors. I noted these insights because they would later support my correlation analysis.

Date: 06-11-2025

Today I setup my work area, got rid of temporary files, and made templates for analysis. I did this mainly for the purposes of clearing but also to get used to the methods by practicing on a smaller dataset first. I did run of visualizing the actual dataset by creating heatmaps on a small sample dataset.

Date: 12-11-2025

Today, worked on the numerical features of the dataset and the analysis of these features. To do this, I produced the correlation heatmaps that allowed me to see the relationships between the features and the target variable. The determining the variables might have more predictive power for the modelling stage of our project.

Date: 17-11-2025

Today I studied about the process of target variable is interacting with several numerical attributes. By comparing the patterns and trends of the groups that defaulted and those that did not, I was able to bring out some significant differences in behavior. From these results, there was a direct influence on the part of understanding the data and prioritizing when creating models from the features that should figure prominently.

Date: 21-11-2025

My focus today was handling outliers. I applied IQR-based capping techniques to extreme values that could negatively impact model stability. After capping, I reviewed each feature distribution to ensure the transformations did not distort the underlying data trends. This step strengthened the reliability of our dataset before entering the modelling phase.

Date: 25-11-2025

I performed the dataset splitting required for modelling. I performed stratified sampling and split the data into training and testing sets, keeping the distribution of the classes as in the original dataset. Besides, I checked that there was no leakage and confirmed that both sets were appropriate for the logistic regression and gradient boosting models scheduled for the coming days.

Date: 28-11-2025

Today I worked on the evaluating the baseline Logistic Regression model. After fitting it on the training data, I measured its performance using accuracy and classification metrics. Although simple, this model allowed us to set a benchmark against which more advanced models could be measured. The baseline results also provided insight into how well linear models capture default patterns.

Date: 29-11-2025

Today I worked on the XGBoost model, main ML algorithms we are using, After configuration, I trained the model and gained its accuracy and predictive in behaviour. The model outperformed logistic regression, confirming the strong value of ensemble boosting for credit default prediction. I saved and documented all results for comparison.

Date: 02-11-2025

I created the confusion matrix and ROC curve for the last selected model. These graphic tools played a significant role in understanding the model's capability to separate the two groups, defaulting and non-defaulting customers. Besides, I delivered a reappraisal of the final decision, and thus, the team got a clear picture of the model's advantages, disadvantages, and applicability in the real world.

Overall Summary

Throughout the duration of this project, I contributed extensively to data analysis, preprocessing, model development, and final evaluation. My work included correlation heatmaps, feature–target relationship analysis, outlier handling, train–test splitting, implementation of baseline Logistic Regression, and building the XGBoost model. I also completed the final confusion matrix and ROC visualisation, supporting the selection of the best predictive model. These contributions played a key role in shaping an accurate and interpretable credit default prediction solution.

Overall percentage contribution to the project: 33.33%