**EDA:**During the data inspection phase, I loaded the dataset from a text file and reviewed the initial rows to understand its structure and contents. This step allowed me to get a sense of the data's layout and identify any potential data issues.

In the data cleaning process, I noticed that some columns had combined values and needed to be parsed correctly. I addressed this issue by extracting individual column values, ensuring data integrity.

Data conversion was crucial, as certain columns were in the wrong data type (object) instead of numeric. I converted these columns to the appropriate numeric data types (float or integer) to facilitate further analysis and modeling.

To understand the data distribution, I created histograms and box plots for numerical variables like 'AGE', 'INCOME', 'LOANS', and 'RISK'. These visualizations helped me identify outliers and skewed patterns.

In the correlation analysis, I calculated the correlation matrix to assess relationships between numerical variables. This step helped me identify potential multicollinearity among features.

I utilized bar charts and pie charts to visualize the distribution of categorical variables like 'MARITAL' and 'RISK', gaining insights into the proportion of different categories within each variable.

Ensuring the balance of the target variable 'RISK' was crucial, as an imbalanced target could lead to biased model predictions.

The heart of the analysis was building and evaluating multiple classification models, including Logistic Regression, KNN, Gradient Boosting Classifier, and SVM. I assessed the models based on accuracy and ROC AUC to measure their performance in identifying reliable customers.

I interpreted the Gradient Boosting Classifier model to understand feature importance, determining which features significantly contributed to the model's predictions. This insight allowed me to identify key factors influencing customer reliability.

Finally, I employed the trained Gradient Boosting Classifier to predict the probabilities of customers being reliable. Using these predictions, I selected the top 5% of customers with the highest probabilities, signifying their reliability.

**Model building:**

After preparing the data, I built and evaluated several classification models, including Logistic Regression, KNN, Gradient Boosting Classifier, and SVM.

The Gradient Boosting Classifier showed the best performance with an accuracy of 72% and an ROC AUC of 78%. Using this model, I identified the top 5% most reliable customers based on their predicted probabilities. Income was found to be the most influential factor in determining reliability, followed by age and loans.

**Conclusion:**

Overall, my analysis revealed valuable insights to help the bank target and promote to these reliable customers effectively. With a focused approach on the top 5% customers, the bank can improve their marketing strategies and enhance customer satisfaction.

**Result:**

Accuracy: **0.72**

ROC AUC: **0.7916666666666667**

**Top 5%**  
 AGE INCOME LOANS RISK\_Prob

206 35 49600 1 0.957730

189 22 17548 1 0.955753

203 45 58381 0 0.953044

188 22 17704 1 0.949656

186 22 17917 1 0.949656

233 35 32827 0 0.945191

235 32 31778 0 0.945191

144 32 31295 0 0.945191

187 25 17886 1 0.940979

213 45 44294 1 0.936365

204 39 52495 1 0.930347

207 39 47161 1 0.930347

**Analysis contribution by features**  
  
