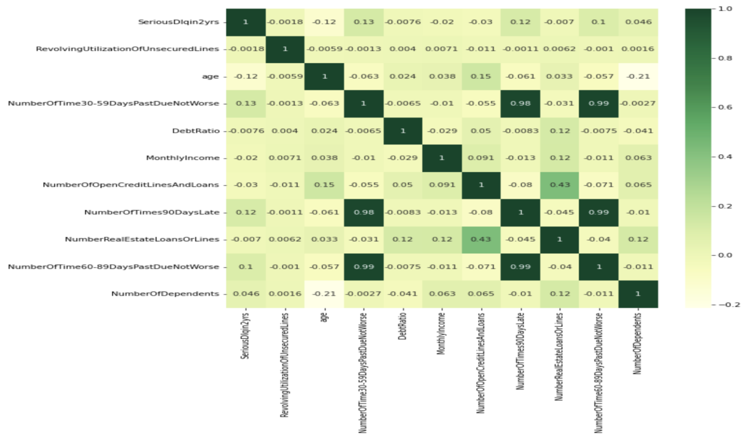
4. Answer the following questions:

For Part 1:

1. What are the factors that have high correlation with the probability of loan default?



From the correlation matrix, we can see thatNumberOfTimes90DaysLate has a correlation value of 0.12, NumberOfTime30-59DaysPastDueNotWorse has a correlation value of 0.13 and NumberOfTime60-89DaysPastDueNotWorse has a correlation value of 0.1. These three independent variables have a high correlation with the probability of loan default.

2. Are there interaction effects occurring among the variables?
From the correlation matrix, we can see that NumberOfTimes90DaysLate, NumberOfTime30-59DaysPastDueNotWorse, and NumberOfTime60-89DaysPastDueNotWorse have a strong correlation.

3. Any other preliminary analysis of the given dataset.

- We did a bar graph for label distribution and found that the class is imbalanced.
- Plotted pie chart based on age and the number of days late. From the pie chart, we can see that the number of times borrowers have been 30 to 59 days or 60 to 89 days, or 90 days or more past due the age group from 25-50 has the highest percentage. For the age group 0 to 25, the number of times the borrower has been 30 to 59 days past due is 15.58% which is less compared to the 60 to 89 days and 90 days or more. It is clear from the graph that the age group from 25 to 50 has the highest percentage of borrowers who have past due.
- Plotted stacked bar graph for finding the relation between Monthly Income and NumberRealEstateLoansOrLines based on SeriousDlqin2yrs. The graph makes it clear that people with low-income levels have mortgage and real estate loans and are frequently in default by 90 days or more.
- Plotted bar graph for analyzing the relation between Age, Monthly Income, and Debt ratio. We can observe that no delinquency has gradually increased and decreased as income and age have increased. The debt ratio is displayed in the graph's line chart. The debt-to-income ratio is slowly rising as people's ages rise. No delinquency and delinquency have almost the same debt ratios.

For part 2:

1. Tell us how you validate your model and why you chose such evaluation technique(s).

- The validation technique used in the model is splitting data into training and validation sets. The reason for doing this is to understand what would happen if the model were faced with data, it has not seen before. In our model, we split the data into a 70% training set and a 30% validation set.
- The model can be improved in the future using k-fold cross-validation. The data is divided into k folds, then trained on k-1 folds, and tested on the one-fold that wasn't included. It performs this for all possible combinations, averaging the outcome for each instance. The fact that all observations are used for validation and training and only once for each is advantageous.
- When using k-fold with grid search for this project, the time required increases. So, in order to validate our model easily, I applied the dataset-splitting strategy.

2. What is AUC? Why do you think AUC was used as the evaluation metric for this challenge?

- A binary classifier's performance is measured by the AUC (Area Under the Curve) score. It shows the likelihood that a randomly selected positive case would be ranked higher than a randomly selected negative instance by the classifier. A classifier that performs no better than random guessing has an AUC value of 0.5, whereas a perfect classifier has an AUC score of 1.0.
- AUC is a suitable option when the classes are not evenly represented in the data since it is not sensitive to class imbalance. That is the reason why AUC is used as the evaluation metric for this challenge

3. What other metrics do you think would also be suitable for this competition?

- A binary classifier's performance is frequently assessed using the metric of **precision**. For binary classification, precision is a useful statistic since it conveys the "exactness" of the classifier.
- **Recall** it helps the classifier be "complete," which is important if the cost of producing a false positive prediction is minimal. When determining the completeness of the classifier, recall is a useful metric to utilize.
- **F1 macro** balances precision and recall while giving the positive and negative classes equal weight, making it a useful statistic for binary classification. As the dataset is imbalanced, it's better to use F1 macro instead of F1.

4. Short explanation of what you tried. What worked and what did not work (ie. you might have tried different features/models before the final one).

- Firstly we tried the logistic regression model which gave 0.50841 accuracy.
- Next, we tried a random forest model which gave accuracy of 0.64156.
- XGBClassifier model gave accuracy of **0.86794**.

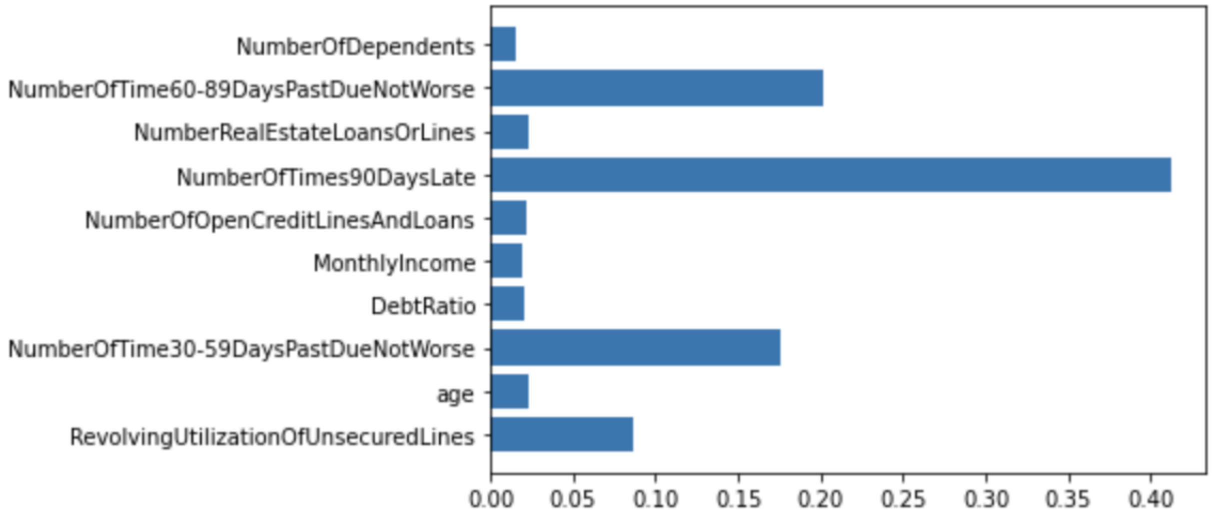
In XGBClassifer we tried dropping different features and below is the accuracy.

Features dropped	**Accuracy**
DebtRatio and NumberOfDependents	0.85981
DebtRatio	0.86606
NumberOfDependents	0.86184
NumberRealEstateLoansOrLines	0.86637
RevolvingUtilizationOfUnsecuredLines	0.84802

Note: We tried manual feature selection but there is a better way to do. I tried dropping one of the highly correlated variables such as NumberOfTimes90DaysLate but it dropped the accuracy.

5. What insight(s) do you have from your model(s)?

From our model, the significance of each feature in the dataset is one of the key insights that can be obtained from the XGBoost classifier. We used XGBoost's built-in function called feature significance to find which feature has the highest importance. From the bar graph below, we can see that the highest score is for the feature NumberOfTimes90DaysLate followed by NumberOfTime60-89DaysPastDueNotWorse, NumberOfTime30-59DaysPastDueNotWorse, and RevolvingUtilizationOfUnsecuredLines. Decision trees are the foundation model of XGBoost. The XGBoost algorithm can be easily modified using hyperparameters to improve performance on our datasets.



6. Can you get into the top 100 of the private leaderboard or even higher?

Yes, the private score for our model is 0.86794. It comes in **position 59.**

To get a higher score we can do k-fold validation which might improve the accuracy. Use feature selection methods like correlation-based feature selection, Recursive feature elimination. Also as we know the class is imbalanced we can try undersampling the majority classes and oversampling the minority class.