Analysis & Planning:

Below is the planning for this assignment. The below steps shall be done using Python or Pyspark wherever necessary.

Data understanding, preparation and EDA

1. Understand the schema of the data and look for following:
    a. Understand the list of columns
    b. Check for any invalid values (values that are unusual for the column)
    c. Identify null values and remove those which does not make any sense, do missing value imputation for those which are needed by proper imputation techniques
    d. Understand the data type of each column and do any necessary conversion
    e. Check for any outliers and remove them which does not cater to this case study
    f. Check duplicates for any data and any type of data redundancies (do this post merging of the 2 dataframes)
2. Build the target variable - The user ID will be classified as delinquent if the customer has ever delayed their payment by more than 60 days.
3. Answer below questions as part of EDA.
    a. What is the proportion of females in the applicant customer base?
    b. Is homeownership higher among male applicants or female applicants?
    c. Is there any correlation between the customer's income level and education level?
    d. What is the average and median salary of the applicant base?
    e. Is the proportion of bad customers higher for people who own cars?
    f. Is the proportion of bad customers higher for those living on rent than the rest of the population?
    g. Is the proportion of bad customers higher for those who are single than married customers?
4. Do visualizations on individual variables and link it to the target variable also. Identify correlations between target and independent variables.
5. Divide the available columns into categorical and numerical data.
6. For each of the categorical column do WOE-IV implementations (use the python code we have) and IV values less than 0.002 to be insignificant in the modelling process

Model building & Evaluation

7. Once above is done and unnecessary categorical columns are removed do the below for the rest of the remaining columns.
    a. Continuous columns with numerical data - No processing required. Here AMT_INCOME_TOTAL, DAYS_BIRTH, DAYS_EMPLOYED,CNT_CHILDREN, CNT_FAM_MEMBERS

    b. Nominal categorical variables represented with integers. - One-hot encoding - here FLAG_WORK_PHONE, FLAG_EMAIL
    c. Ordinal Categorical variables with integer representation - No processing required
    d. Nominal Categorical variables represented with strings - String indexer + One hot encoding. Here NAME_INCOME_TYPE,NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, CODE_GENDER, FLAG_OWN_CAR,FLAG_OWN_REALTY, OCCUPATION_TYPE
    e. Ordinal categorical variables represented with strings - String indexer. (make sure the right categories get the right integer)

8. Use a train-test split of 70:30 and a seed value of '2018' in your PySpark model. Refer https://spark.apache.org/docs/latest/api/python/pyspark.sql.html#pyspark.sql.DataFrame.randomSplit
9. Build Model 1 with all the remaining available columns and evaluate below metrics
    a. KS statistic - identify probability threshold
    b. AUC
    c. Confusion Matrix (Precision, recall, F1-score) using above probability threshold
10. Since we need to target for a higher recall let us do some model tuning by checking the VIF values for available columns and remove those with high VIF (also consider business aspects) using statsmodel.
11. Construct Model 2 and again check AUC and Confusion matrix with probability threshold identified using KS statistic.
12. Using statsmodel identify features with low variance (variance thresholding) and see if anything can be removed and Model 3 can be constructed.
13. Use Select K best for feature selecting. See if https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html can be used for this.
14. Also check if multiple columns can be combined together to do any feature transformation. Check CTR assignment.
15. Check the p values of the columns. See if https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression can be used similar to done for Linear Regression.
16. Construct the final model or decide the best model from one of the above based on high recall.

Interpret the model properly and explain the model at each evaluation step before proceeding to next model tuning.

Check if the code is as per coding guidelines in each step.

Implementation:

Calculation of target variable:

Considering that the number of days count was done properly for each month (i.e) added up from previous months, we can take the STATUS column alone for building target variable and leave the MONTHS_BALANCE column from calculation.

If the STATUS value is greater than 1 (more than 60 days) then DELINQUENT should be 1

However if the customer has not got any loan for that month then that record can be dropped as it does not serve our purpose

In this assignment a value of 1 is considered as bad (DELINQUENT) and a value of 0 is considered a good customer.

Model 1 Interpretation:

Precision for bad event (for customer to be DELINQUENT = 1) = 1.0

Recall for bad event (for customer to be DELINQUENT = 1) = 0.01

The above model DELINQUENT cases the recall value is very less 0.01 with a default threshold of 0.5. Hence we can conclude the total available positive data points only few were predicted correctly. Hence this is NOT a good model for our use case.

AUC = 0.6842

F1 Score = 0.0198

The F1 score is less since Recall for the label is less. We need to improve Recall and F1 score.

Probability Threshold = 0.4373 (as per KS Statistic). We shall use this threshold in the 2nd model.

Model 2 Building and Interpretation:

For the 2nd model, we will do identification of multicollinearity to eliminate those features which can easily be identified by other remaining features/independent variables. Having high multicollinearity in a model lowers the scoring of the model. Hence let us calculate the Variance Inflation Factor (VIF) of each feature and eliminate those greater than 10.

Once the features which are dependent on other features are identified, we will drop them and also use the above calculated threshold of 0.4373 for building the 2nd model.

Model 2 Interpretation:

Precision for bad event (for customer to be DELINQUENT = 1) = 0.66

Recall for bad event (for customer to be DELINQUENT = 1) = 0.02

AUC = 0.6731

F1 Score = 0.0388

We can see that Recall has improved as a result Precision has declined. However there is good improvement in F1 score. Though the area under the ROC curve has slightly decreased it is not much. Since we are more concerned about Recall **Model 2 is better than Model 1.**

NOTE: We have used threshold of 0.4373 identified using KS Statistic and will continue using the same instead of default which is 0.5

Model 3 Building and Interpretation:

For Model 3 let us use the GLM algorithm in statsmodel and check the significance of each of the independent variables. Here we will check which independent variable has a high chance of having their coefficient to be close to 0 (hence insignificant as per null hypothesis) and eliminate that.

Model 3 Interpretation:

Precision for bad event (for customer to be DELINQUENT = 1) = 0.66 (same as previous model)

Recall for bad event (for customer to be DELINQUENT = 1) = 0.02 (same as previous model)

AUC = 0.6555 (decreased but slightly only)

F1 Score = 0.0388 (same as model 2)

Though there is a decrease in AUC, other metrics which are important in our case as Recall and F1 score are the same even after eliminating many features.

Hence Model 3 is better than Model 2 has there are less features and hence simpler than Model 2.

Currently in the above Model below are the final list of features used from the source dataset.

- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- CODE_GENDER
- FLAG_OWN_REALTY
- OCCUPATION_TYPE
- AMT_INCOME_TOTAL

- DAYS_EMPLOYED

Though we can further look to eliminate some features, we are deciding to keep them since they might have business importance (like CODE_GENDER, OCCUPATION_TYPE, NAME_INCOME_TYPE)

Though the p value for AMT_INCOME_TOTAL is high and ideally should be eliminated, we are keeping it as it is the only numerical value which gives the source of income for our customers.

Final Interpretation:

From the set of features we can say that none of the features are really strong predictors in this case. We should look to get more features which decides whether a customer is DELINQUENT or not. However with the available data we can say the last model (**Model 3)** is the best possible since we have eliminated a lot of features at same time and also improved Recall from Model 1.

Conclusion: Model 3 is the best possible model with the given list of features and data with class imbalance.