**Task-2 :**

**Explain your intuition behind the features used for modeling**.

Ans- I have removed applicant ID and analysed the rest of the data.Reason for removal of applicant ID is that value is unique and it is similar to loan application ID.

So after joining the 2 csv files it is not required to keep 2 unique values.

**Are there missing values? If yes how you plan to handle it.**

Ans– Yes there are lot of missing values in some features.

There are different type of imputation techniques such as {mean,median,mode} replacement, model based imputation etc.

Ex– If the data is categorical then it s converted to numerical first, then if there is a outlier median based imputation is followed else mean based method is followed

* Sometimes most repeated value in the column is replaced i.e mode based imputation.

In this assignment I have followed mode based imputation such that idea here is most repeated value in the column has higher probability of being present in the empty cell.

**How categorical features are handled for modeling.**

Ans-

***1)Numerical category:***

Here data is usually left as it is, since the model can understand integer or float values.

***2)String category:***

Here mainly 2 types are there ordinal encoding & one hot encoding.

[ **Note:** Have to be careful while doing ordinal encoding because there should be some order in a feature category to apply this method.

Ex– v.good,good,bad,worst]

One hot encoding– preferred generally in linear models

Ordinal – preferred generally in tree based models

**Describe the features correlation using correlation matrix. Tell us about few correlated feature & share your understanding on why they are correlated.**

Ans-

| **applicant\_id** | **Months\_loan\_taken\_for** | **Principal\_loan\_amount** | **EMI\_rate\_in\_percentage\_of\_disposable\_income** | **Has\_coapplicant** | **Has\_guarantor** | **Number\_of\_existing\_loans\_at\_this\_bank** | **Primary\_applicant\_age\_in\_years** | **Number\_of\_dependents** | **Years\_at\_current\_residence** | **Foreign\_worker** | **high\_risk\_applicant** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **applicant\_id** | 1 | 0.009359 | -0.056669 | 0.002269 | -0.012961 | 0.01186 | -0.038409 | -0.010583 | 0.017048 | -0.028017 | 0.066389 |
| **Months\_loan\_taken\_for** | 0.009359 | 1 | 0.624984 | 0.074749 | 0.029698 | -0.039594 | -0.011284 | -0.036136 | -0.023834 | 0.034067 | 0.138196 |
| **Principal\_loan\_amount** | -0.056669 | 0.624984 | 1 | -0.271316 | 0.079076 | -0.065237 | 0.020795 | 0.032716 | 0.017142 | 0.028926 | 0.05005 |
| **EMI\_rate\_in\_percentage\_of\_disposable\_income** | 0.002269 | 0.074749 | -0.271316 | 1 | -0.013048 | -0.006429 | 0.021669 | 0.058266 | -0.071207 | 0.049302 | 0.090024 |
| **Has\_coapplicant** | -0.012961 | 0.029698 | 0.079076 | -0.013048 | 1 | -0.048426 | -0.006001 | -0.018357 | -0.032817 | 0.001623 | -0.066338 |
| **Has\_guarantor** | 0.01186 | -0.039594 | -0.065237 | -0.006429 | -0.048426 | 1 | -0.024682 | -0.023923 | 0.036589 | -0.028334 | -0.097256 |
| **Number\_of\_existing\_loans\_at\_this\_bank** | -0.038409 | -0.011284 | 0.020795 | 0.021669 | -0.006001 | -0.024682 | 1 | 0.149254 | 0.109667 | 0.089625 | 0.009717 |
| **Primary\_applicant\_age\_in\_years** | -0.010583 | -0.036136 | 0.032716 | 0.058266 | -0.018357 | -0.023923 | 0.149254 | 1 | 0.118201 | 0.266419 | 0.006151 |
| **Number\_of\_dependents** | 0.017048 | -0.023834 | 0.017142 | -0.071207 | -0.032817 | 0.036589 | 0.109667 | 0.118201 | 1 | 0.042643 | -0.077071 |
| **Years\_at\_current\_residence** | -0.028017 | 0.034067 | 0.028926 | 0.049302 | 0.001623 | -0.028334 | 0.089625 | 0.266419 | 0.042643 | 1 | 0.054097 |
| **Foreign\_worker** | 0.066389 | 0.138196 | 0.05005 | 0.090024 | -0.066338 | -0.097256 | 0.009717 | 0.006151 | -0.077071 | 0.054097 | 1 |
| **high\_risk\_applicant** | -0.029125 | 0.214927 | 0.154739 | 0.072404 | 0.062728 | -0.055039 | -0.045732 | -0.091127 | -0.003015 | 0.002967 | 0.082079 |

* From the above sheet we can observe that principal loan amount and months loan taken for have high correlation of 0.6 ( it is the highest among all).

Reason for this is if someone taken larger principal sum of loan amount they tend to have larger duration to payback.

**Do you plan to drop the correlated feature? If yes then how.**

Ans–

No , I didn’t remove correlated features because it depends on the algorithm which we use. If there is lot of correlations then its better to remove due to Curse of dimensionality and it will have an effect on algorithms like KNN(distance based).

Since I haven’t used KNN I didn’t drop the correlated features .

**Which ML algorithm you plan to use for modeling.**

Ans– I have decided to use logistic regression , Support vector machines and Random forest.

**How you will select the hyperparameters for models trained in above step.**

Ans–

* First I will train on default values and then model is trained on cross-validation of data.If the metric didn’t change significantly
* Model is to be tested with different hyperparameters like lambda(or C) in linear models , depth in tree models and depth+ sampling in ensembles like GBDT or Random forest.

**Which metric(s) you will choose to select between the set of models.**

Ans–

For classification tasks there are different metrics such as :

1)ROC-AUC

2)ACCURACY

3)F1-SCORE

4)LOG-LOSS

Among these F1-score and ROC -AUC is best for imbalanced datasets(it covers all aspects of the data such as TP,FP,TN,FN) and accuracy is worst for imbalanced dataset.

***I have chosen F1-score as main metric to evaluate models.***

**Explain how you will export the trained models & deploy it for prediction in production.**

Ans– Currently learning so don’t have proper idea on it, (I will learn quickly and till now I am concentrating more on classical ML and DL)