REUNION ASSIGNMENT

By: Vibhuti Dabas

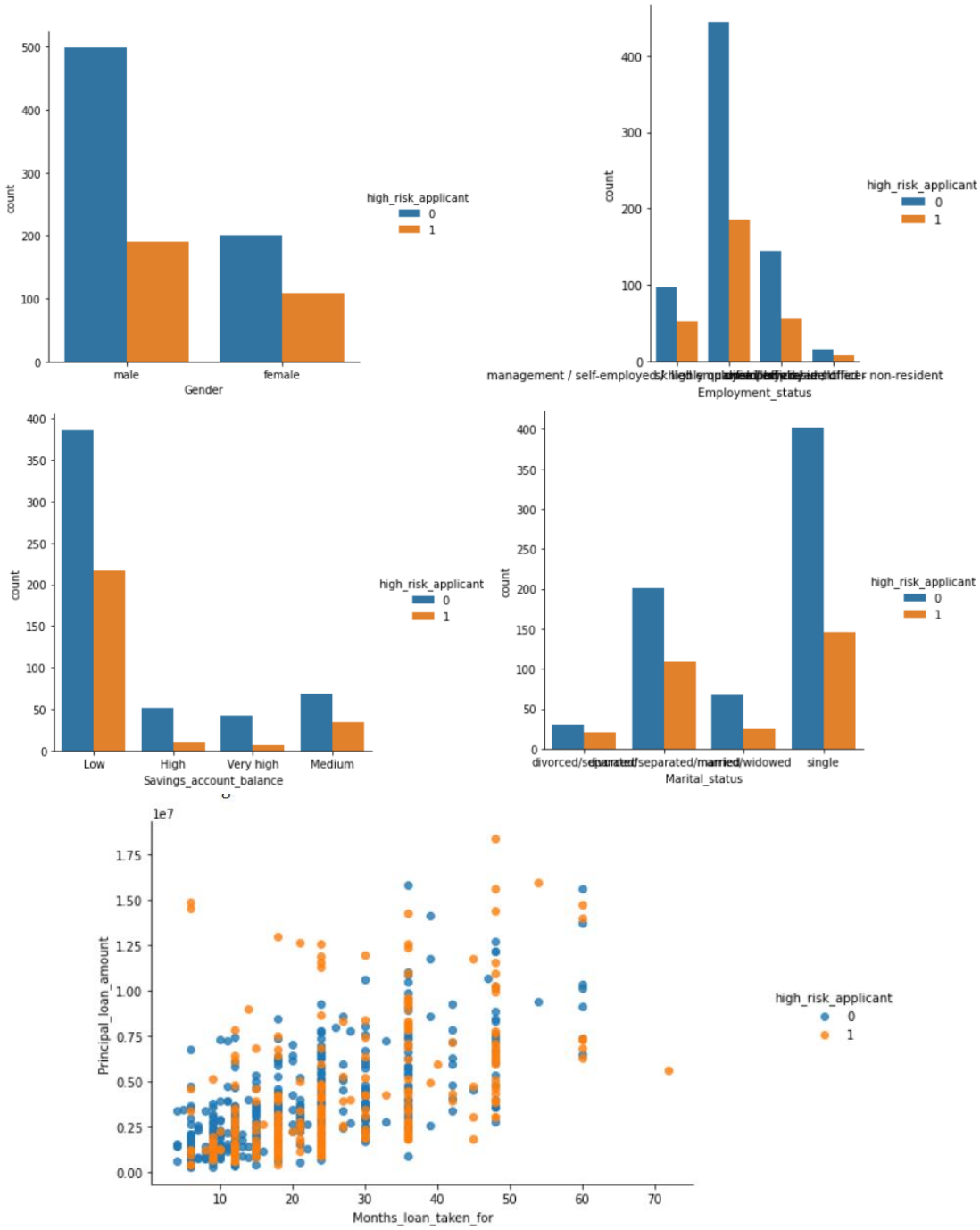
TASK 1

1. **Do the Exploratory Data Analysis & share the insights.**

   After downloading the data application.csv had more information about the applicant. By further exploring the data, found few relevance between attributes.
   Before looking at the relation I removed irrelevant columns like telephone_registered and foreign_worker. Subsequently, checking for columns having NULL values and removing those columns that had more than 30% of NULL values.

2. **How would you segment customers based on their risk (of default).**

   When I explored the data further, I tried to segment the applicants and find patterns by sorting the high_risk_applicant column. It seems that gender, employment_status, savings
   and marital_status had relations. To visualize I used seaborn to plot catplots.



3. **Which of these segments / sub-segments would you propose be approved?**
   For e.g. Would a person with critical credit history be more creditworthy? Are young people more creditworthy? Would a person with more credit accounts be more creditworthy?

   After observing these segments I concluded, difference can be seen in gender, savings and employment, hence dropping the amount_taken.

   Other than these attributes in loan.csv "Property" can be an important factor as it tells you about the financial strength of the applicant.

   A person with more credit-score or more collateral can be creditworthy. Young people can be less creditworthy as their jobs are unstable and don't own many potential

4. **Tell us what your observations were on the data itself (completeness, skews).**

   The data was sufficient but the size was small, due to which accuracy after a certain level could not be achieved. After splitting the data, the training was performed was only on 800 values.

   There could be credit-score as it determines the person's financial history, the data was vague and the columns had many null values and had to remove.

TASK-2

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

   Intuition behind choosing the features was to have relevant attributes that show applicant history and his likelihood of returning the loan. As the telephone or maximum money he has earned does not play any role in determining this. They were removed.

2. **Are you creating new derived features? If yes explain the intuition behind them.**

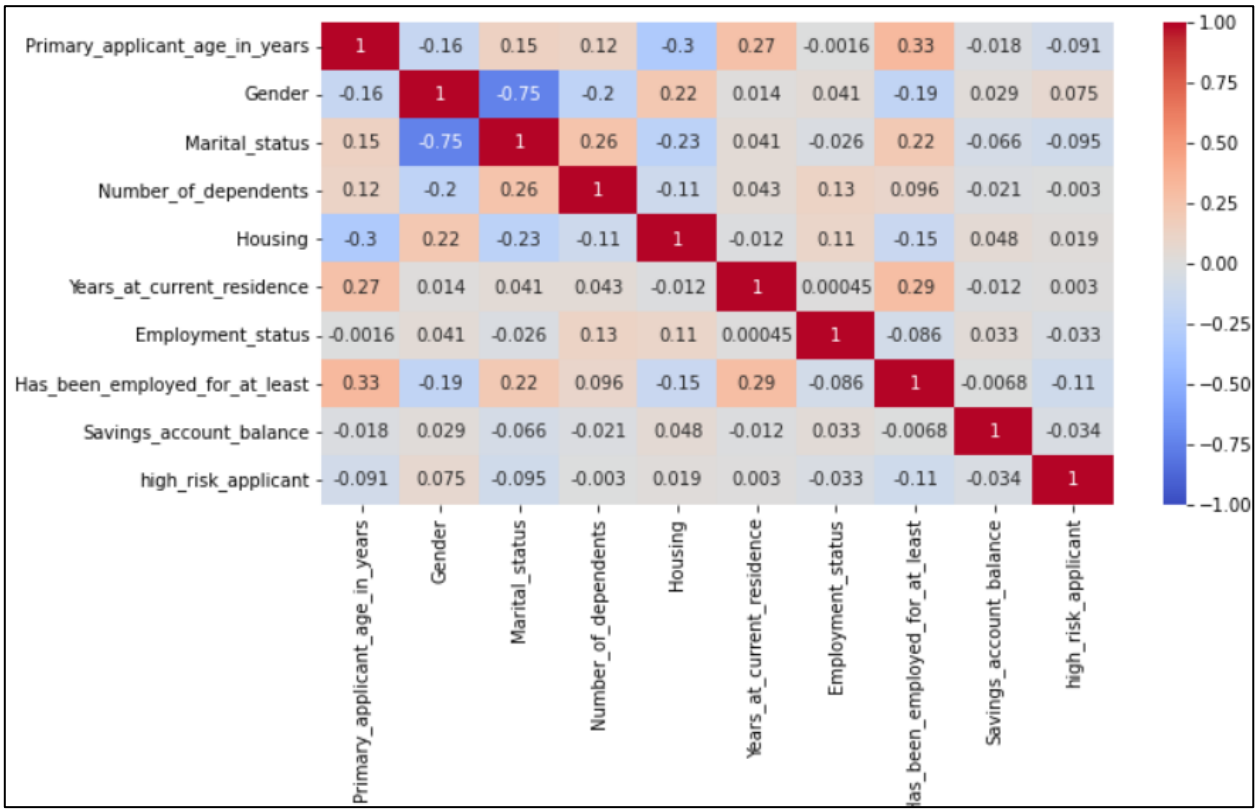
   No, I didn't derive any new features.

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

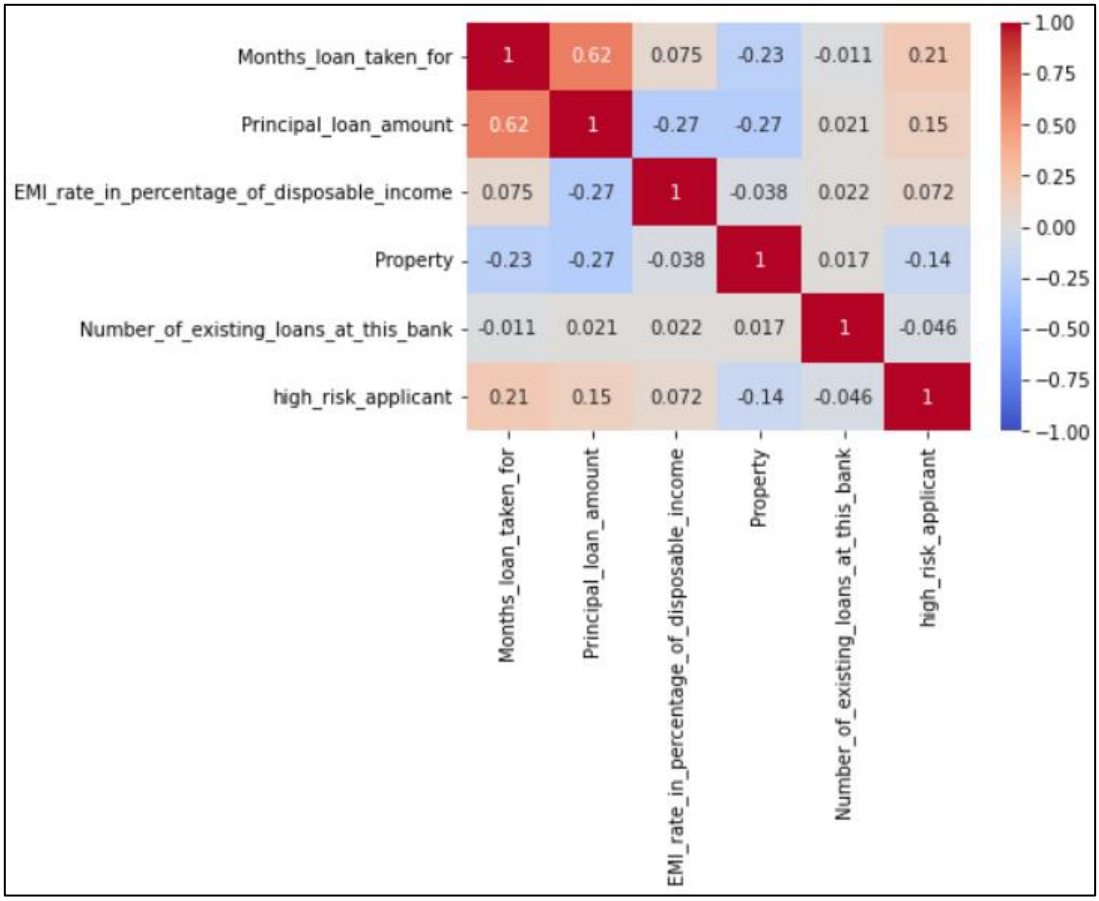
   There were missing values in few columns, few of which were important like savings or least_employement duration.
   Whereas, attributes like lower_limit of balance and upper_limit of balance in the account were not very important. These columns also had NULL values in more than 30% of the column, hence removing them was harmless.

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

Correlation matrix of application.csv is:



Correlation matrix of loan.csv is:



As predicted or observed, the correlation matrix did not find a good relation between the attributes. In application.csv it can be seen that “has_been_employed_for_atleast” had many positive relations as compared to others.
In loan.csv, “Months_loan_taken_for” had a strong relation with “principle_loan_amount” with 0.6 scale and somewhat related to “high_risk_applicant with 0.2 scale.

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

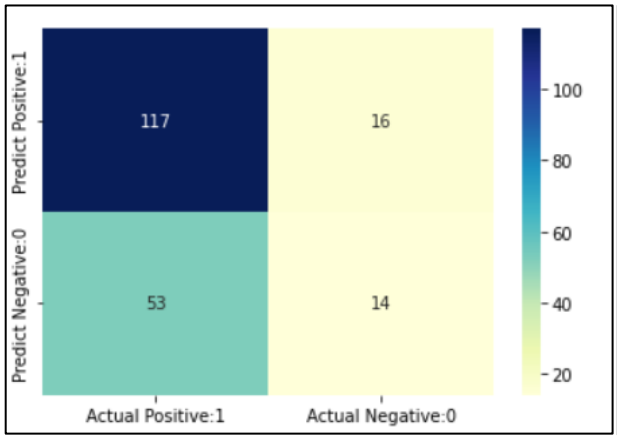
It is not necessary to drop the correlated features/variables, unless correlation is 1 or -1 in which case one of the variables is redundant.

6. **Which ML algorithm you plan to use for modelling.**

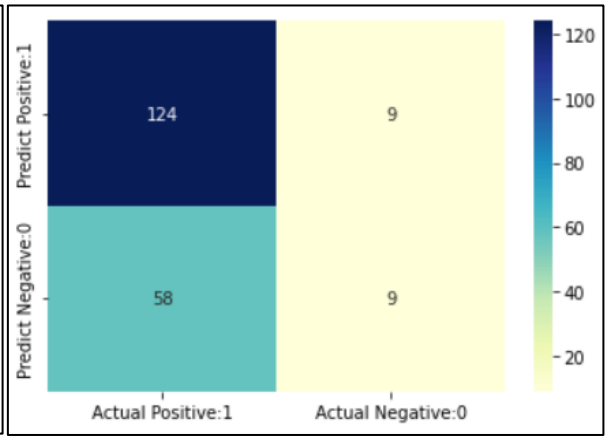
I plan to use 4 different binary classifiers:
>Naïve Bayes
>K-Nearest Neighbours
>Random Forest
>Logistic Regression

7. **Train two (at least) ML models to predict the credit risk & provide the confusion matrix for each model.**

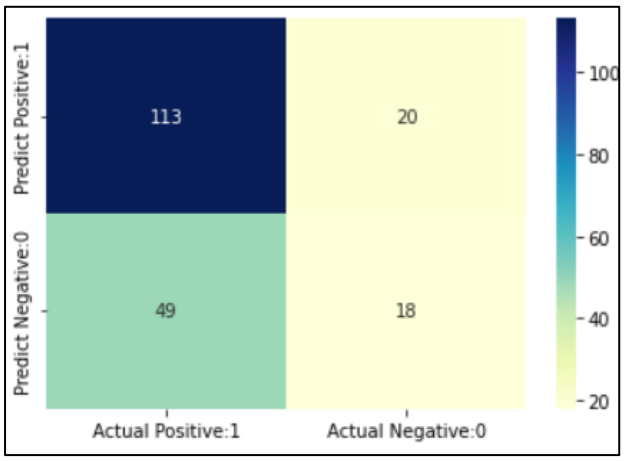
Naïve Bayes:



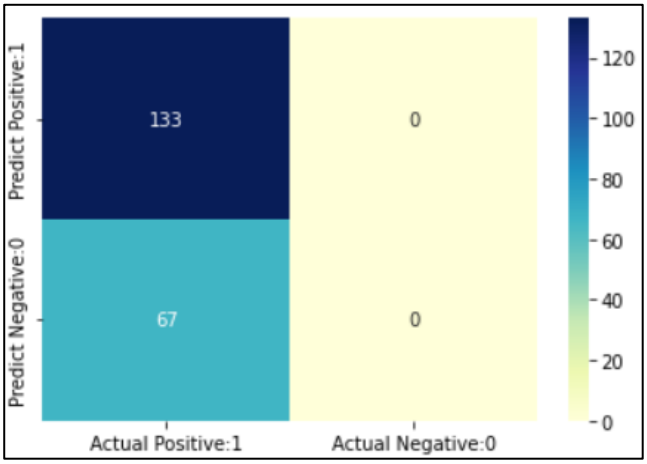
K-NN:



Random Forest:



Logistic Regression:



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

I used pickle library to save models in the colaboratory file and downloaded them from the console on the left.

Saving model

```
pickle.dump(lr, open('log_reg.pkl', 'wb'))          #logistic model
pickle.dump(knn, open('knn.pkl', 'wb'))             #knn model
pickle.dump(gnb, open('naive_bayes.pkl', 'wb'))     #naive bayes model
pickle.dump(rfc, open('random_forest.pkl', 'wb'))   #random forest model
```