DEFAULT CREDIT CARD CLIENT

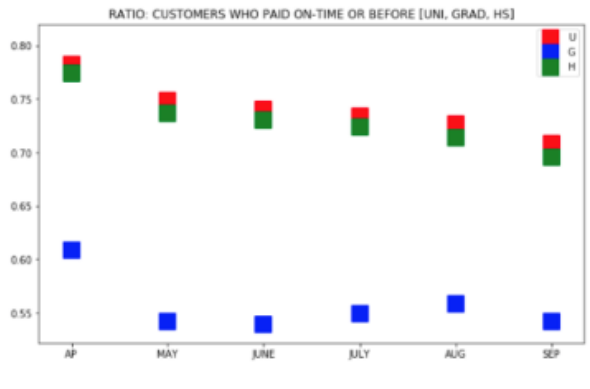
This project is aimed to analyze Default credit card client data set to predict whether a customer will be default or not based on his/her past six months pay records as well as other attributes. Default credit card client data set stored in UCI Machine Learning repository system consists of 24 attributes and 30,000 instances. Various machine learning (ML) algorithms such as Logistic Regression, GaussianNB, Decision Tree Classifier, Extra tree Classifier, have been applied to predict the default by predicting the accuracy of the ML model and the accuracy is very good. Details analysis of the data shows that based on features selection of the independent variables, the prediction accuracy turns out to be too excellent with some of those ML models such as Random Forest Classifier, Extra Trees Classifier or Ensemble Gradient Boosting Classifier. Also it shows that the ratio of customer who pays on time drops down with time, university and high school graduates pay records are better grad students and single customer are more creditable than married group.

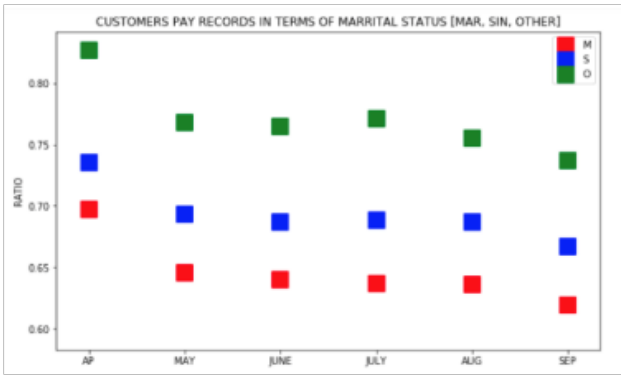
Data Visualization

Among the following twenty four attributes such as 'ID', 'LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT2','BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1','PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6' , analysis was carried over based on based on selected features.

As we analyze the six pay cycle we found the number of clients who paid on time or before the time decreases with pay cycles as shown here.

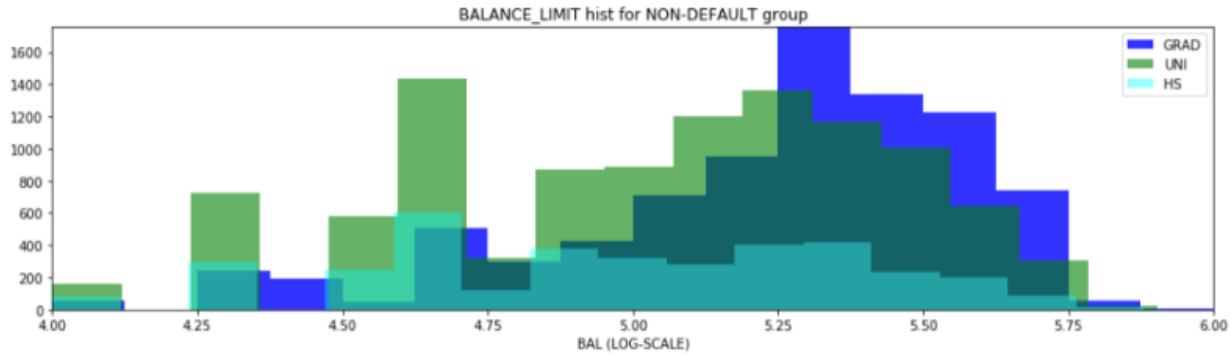
However, in terms of other attributes such as education, marital status we found that university and high school grads pay records make consistent difference comparing to graduate students.





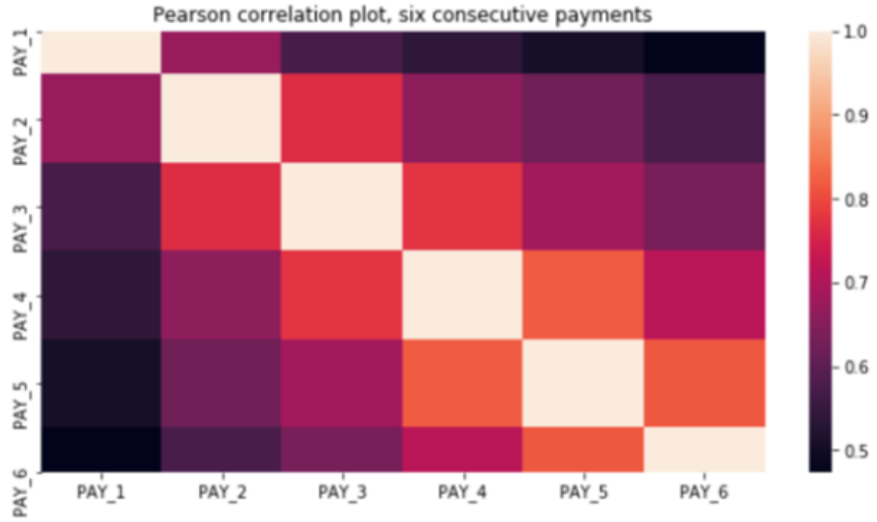
In terms of marital status single and other groups pay records are better than married group. In terms of age group, we found 27-32 years age group is the largest among clients irrespective of default or not default status.

In terms of credit limit, the largest group of customers credit ranges from a tenth of a million to three quarters of a million, but the density of grads mostly populates towards high credit limit end in the hist.



FEATURE CORRELATION:

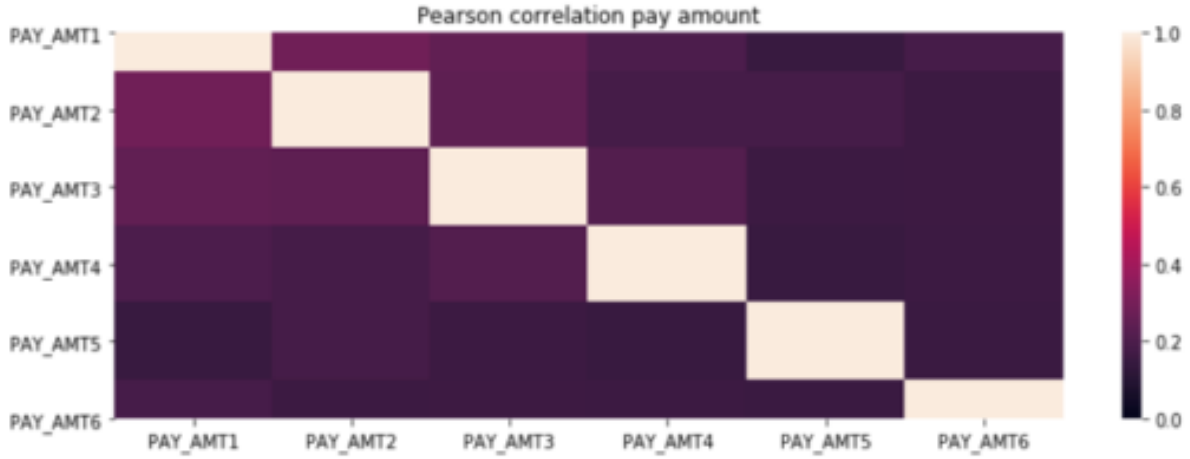
We found consecutive pays (whether paid on time or not) are more correlated (items of correlation matrix ranges from 1 to 0.5)



rather than consecutive pay amount or the ratio of pay amount to bill amount.

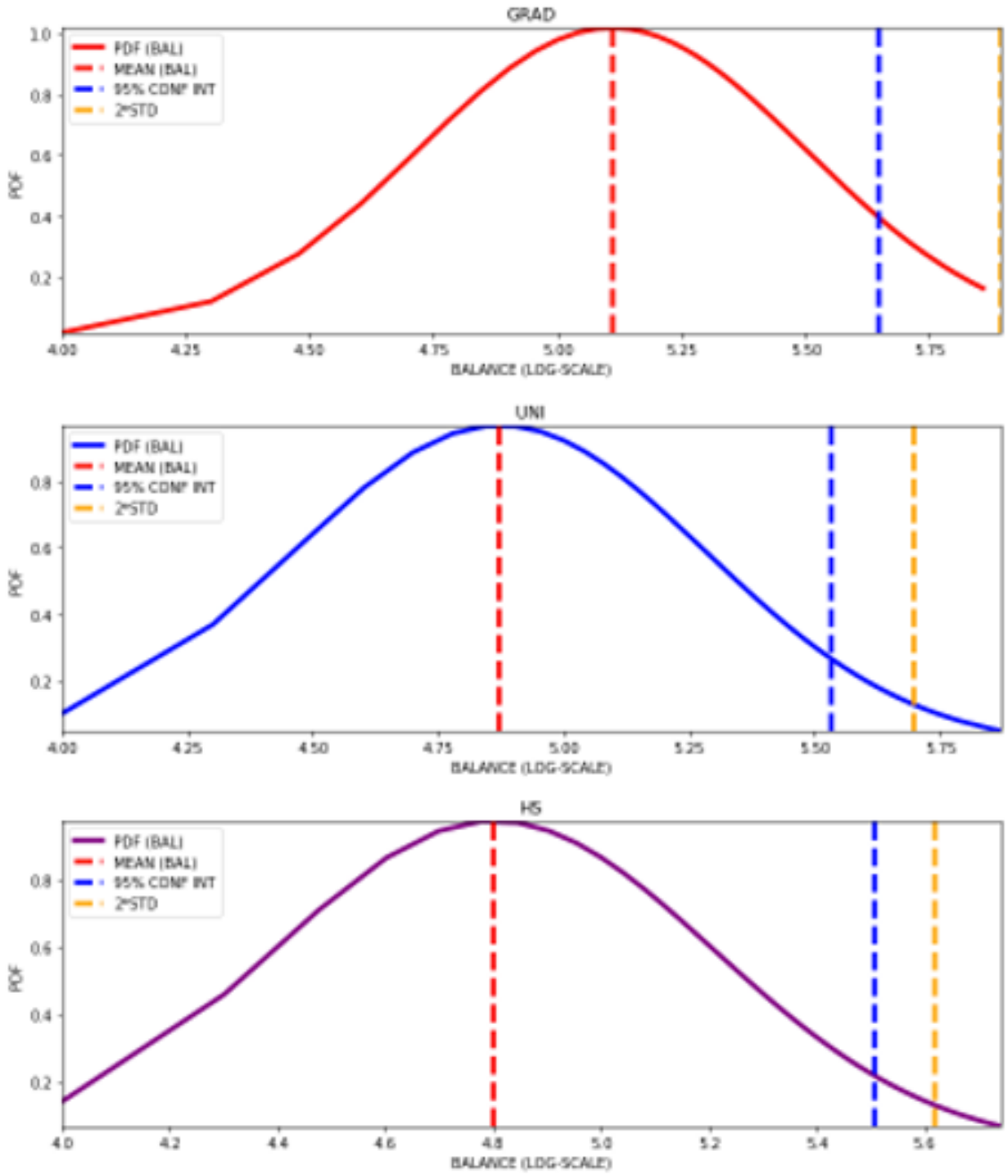
	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
PAY_AMT1	1.000000	0.285576	0.252191	0.199558	0.148459	0.185735
PAY_AMT2	0.285576	1.000000	0.244770	0.180107	0.180908	0.157634
PAY_AMT3	0.252191	0.244770	1.000000	0.216325	0.159214	0.162740
PAY_AMT4	0.199558	0.180107	0.216325	1.000000	0.151830	0.157834
PAY_AMT5	0.148459	0.180908	0.159214	0.151830	1.000000	0.154896
PAY_AMT6	0.185735	0.157634	0.162740	0.157834	0.154896	1.000000

```
<function matplotlib.pyplot.show(*args, **kw)>
```



STATISTICAL ANALYSIS

We found over all clients credit card balance or clients grouped in terms of various attributes such as default, not default group, grouped education, marital status follows similar distribution where the 95% confidence interval with respect to the mean is around the two times the STD of the distribution.



Though the distribution of the three groups are similar, however statistically those groups are not identical as we see from the ttest results.

```
df_GR = df[df['EDUCATION']==1];df_U = df[df['EDUCATION']==2];df_HS = df[df
['EDUCATION']==3]
ttest_ind(df_GR['BAL'].values, df_U['BAL'].values, equal_var=False),ttest_
ind(df_GR['BAL'].values, df_HS['BAL'].values, equal_var=False), ttest_ind
(df_U['BAL'].values, df_HS['BAL'].values, equal_var=False)
```

```
(Ttest_indResult(statistic=39.57867154935797, pvalue=0.0),
 Ttest_indResult(statistic=41.3050193831098, pvalue=0.0),
 Ttest_indResult(statistic=10.692676069537066, pvalue=1.5857820246089056e-
26))
```

since the p values are less than 0.05.

MACHINE LEARNING PREDICTIONS:

We applied various ML algorithms to predict the accuracy of the model. Cross validation among the model shows Decision Tree Classifier, Extra Trees Classifier and Random Forest Classifier predict excellent ROC square based on feature selection. Also, the correlation matrix revels that certain groups are more correlated than others, so this result is expected.

```
MLclf = [LogisticRegression(), GaussianNB(), DecisionTreeClassifier(), ExtraTreesClassifier(), RandomForestClassifier
roc_list = []
for clf in MLclf:
    y_p = clf.fit(x_tr, y_tr).predict(x_t)
    roc_list.append(roc_auc_score(y_t, y_p))
    print(roc_auc_score(y_t, y_p))
```

```
0.754805224490905
0.861962574825473
0.9983542841660619
0.9918186753431147
0.9825125334353151
0.9522269863029847
```

BUSINESS VALUE

ML cross validations, grid search, feature selection is very important in terms of investment, economic and business point of view. While the world economy is devastated by pandemic, predictions with higher accuracy is extremely important.

The total investment among the thirty thousand customer is roughly five billion dollar and so improving the accuracy even by 0.001 lead to a net savings of (5 billion) $\times$ (0.001) = 5 million. So this analysis is important from business and economic point of view.