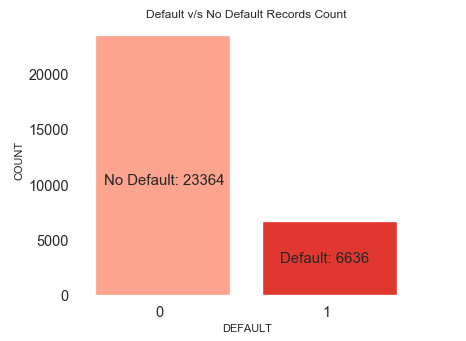
Credit Limit Prediction

## Introduction

This project aims to build a data driven approach to find better ways to understand how much credit to allow someone to use or, at the very least, if someone should be approved or not. Credit approval company has gathered information on 30000 customers. The dataset contains information on 24 variables, including demographic factors, history of timeliness of payments, billed and paid amounts for customers from April 2005 to September 2005, as well as information on the outcome: did the customer default or not?

A quick analysis of data shows that out of 30000 clients, 6636 clients (22.12%) defaulted on their credit balance. There seems to be high number of defaults so it is important to understand if a person should be approved and how much credit limit should be allowed.



## Data Pre-Processing

Some of the categorical variables are in form of labels instead of numeric values. So we converted them to numeric values for building our prediction models.

Education: ['university' 'graduate school' 'high school' 'other']

SEX: ['female' 'male']

Total number records with Education category of ‘other’ are just 468. Similarly, only 54 records with MARRIAGE category of ‘0’. As these categories don’t have significant data, we dropped records with these categories to simplify the analysis and reduce the number of features

Also renamed the column ‘default payment next month’ to just ‘DEFAULT’ as it is makes code more readable. The values for this column are converted into numerical values. So ‘not default’ is mapped to 0 and default is mapped to 1.

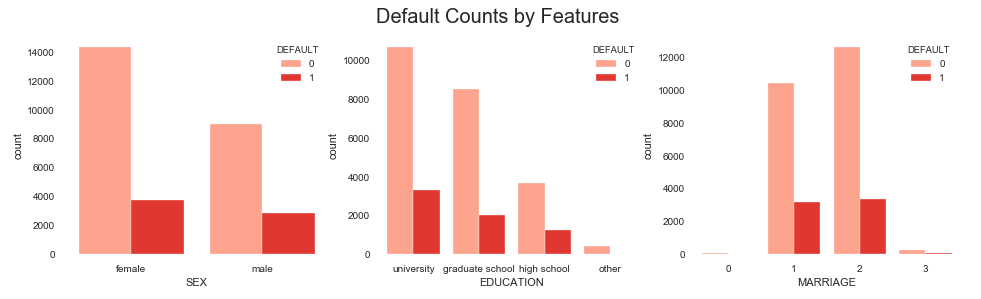
The following repayment status (PAY\_0 to PAY\_6 columns) values can be mapped to 0 (zero) as this reduces the noise in repayment columns. However, payment delay months values of 1 to 8 are more important in predicting defaulters.

*-2: No consumption; -1: Paid in full;*

are mapped to *0: The use of revolving credit;*

## Exploratory Data Analysis (EDA)

This section discusses the finding of exploratory data analysis and these finding also become input to our feature selection process. Following figure shows record counts for Default v/s No Default by different features.

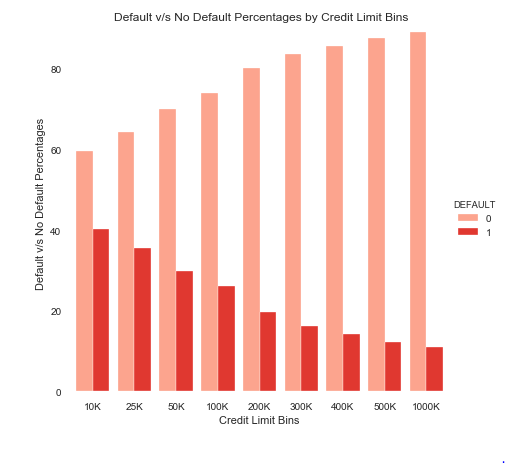
  
The table below represents the defaults in terms of percentages.

|  |  |  |
| --- | --- | --- |
| Gender | Education | Marital Status |
| female 0 79.22%  1 20.78%  male 0 75.83%  1 24.18% | grad school 0 80.76%  1 19.23%  high school 0 74.84%  1 25.16%  university 0 76.27%  1 23.74% | 1(married) 0 76.53%  1 23.47%  2(single) 0 79.07%  1 20.93%  3(divorced)0 74.00%  1 26.01% |

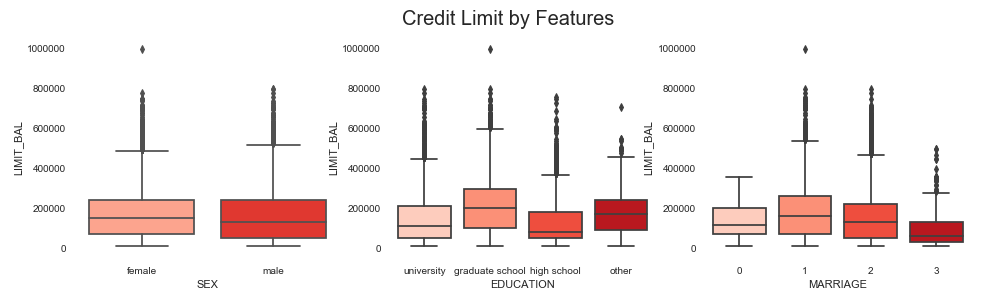
From the data in above table we can infer that:

* Females have slightly lower percentage of defaults than males
* People with graduate degree have slightly lower percentage of defaults than other education levels
* Singles have slightly lower percentage of defaults than married or divorced.

Next we wanted to see if credit limit has any relationship with Defaults. We broke down the credit limit into easy to understand bins and checked if there are more defaults in certain credit limit bins. Following picture shows Default v/s No Default percentage by different bins. It is evident from this graphs that people with higher credit limits have lower percentage of defaults than the people with lower credit limit.



We analyzed the credit limits by various demographic features and noticed that people with Graduate degree have slightly higher credit limits. Similarly, married people have slightly higher credit limits than other marital statuses.



## Modeling

One of the objective of this study is if we can predict the credit limit for a person based on demographics and repayment data. We used 3 different regression models, performed cross validation, computed their accuracy score (R-squared score).

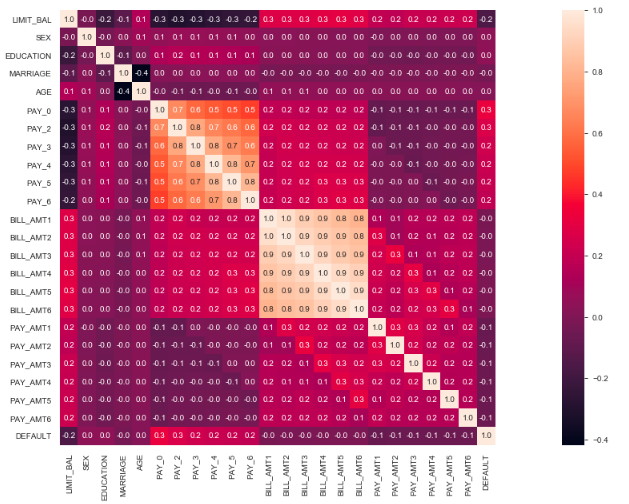
Random Forest Regressor 0.45

Linear Regressor 0.29

Support Vector Reg - 0.052

## Feature Engineering

In order to make our models efficient and avoid any over fitting, we used correlation matrix and used feature engineering (SelectKBest, Recursive Feature Elimination and RidgeRegression), to identify most important 15 features.



Here is the list of features we selected for implementing our models.

'AGE','SEX’,'BILL\_AMT1','BILL\_AMT2','BILL\_AMT3','BILL\_AMT4','BILL\_AMT5','BILL\_AMT6','PAY\_AMT1','PAY\_AMT2','PAY\_AMT3','PAY\_AMT4','PAY\_AMT5','PAY\_AMT6','EDUCATION’

## Model Hyperparameter Tuning

We performed hyperparamter tuning on our selected models to improve accuracy of predictions. The r2 score for RandomForestRegressor improved from .46 to .48

## Results

We calculated R-squared and RMSE scores to check the accuracy of our models.

R-squared ranges from zero to one, with ‘0’ indicating that the proposed model does not improve prediction over the mean model, and ‘1’ indicating perfect prediction.

We got a R-squared of 0.45 which means that model explains only 45% of the variance and it is not a very good fit. We got a high value of approx. 94,000 for root-mean-squared error (RMSE), this means that our model’s prediction could be off by 94,000 from the actual credit limit in our samples.

## Summary

Based on our analysis, we see that it is not possible to predict the credit limit using regression models. These models do not consider the defaults and delayed payments when determining the credit limits.

We also observed that persons in Age group 25 to 45 years, Females, Singles and persons with Graduate degree have a slightly lower percentage of defaults. Persons which have higher credit limits in the existing samples also have lower defaults. The attributes of this population with higher credit limits needs to be studied further to understand which other factors might be contributing to lower defaults. A credit scoring based model which also takes into persons defaults and delayed payment history in addition to other factors could be an ideal solution for predicting credit limit.