**EDA – Loan Performance**

Frank Zhang 06/28/17

Now we have been given two datasets. One includes information of customers who have been given a loan in a 6 month period. The other contains every loan that has been given in this time and whether it has been a good loan or a bad loan. Our goal is to identify which new applicants should be given a loan in the future.

**First Step - Combine Two Datasets Using SQL**

1. Save two datasets into csv files separetely
2. Change format of variable “idLoan” in csv file by using “lower” and “mid” function in Excel
3. Create new feature “Loans\_Num” by indicating number of loans each customer has
4. Import two files into MySQL Workbench
5. Use inner join to combine two datasets by customer\_id

ELECT 
FROM 
Loan.' application data' 
JOIN 
Loan. loan 
ON 
Loan.' application dat

**Second Step – Clean the Data Using Pandas in Python**

1. Read combined csv file into python by pandas

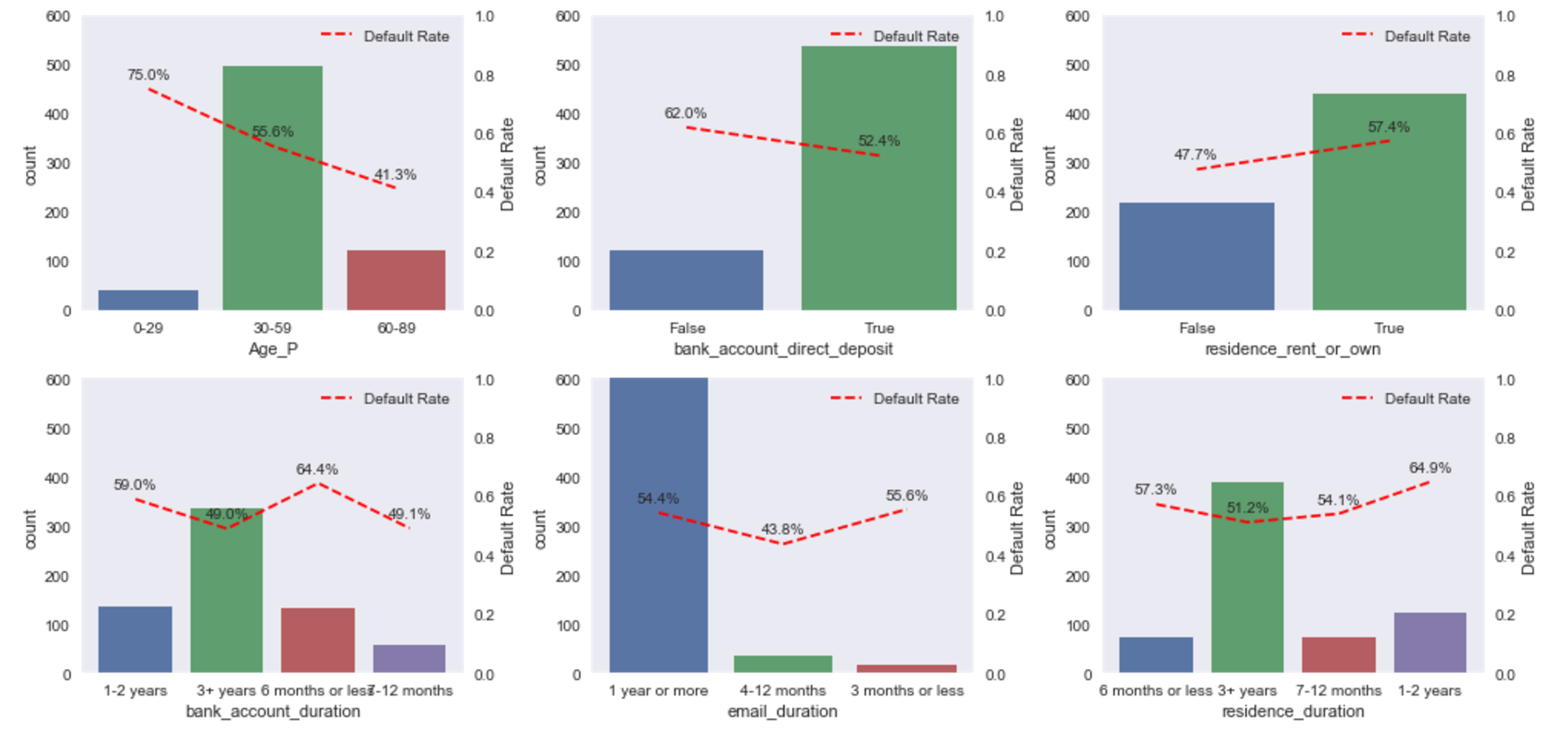
659 observations and 34 features.

1. Create new Features and Fill Missing Values

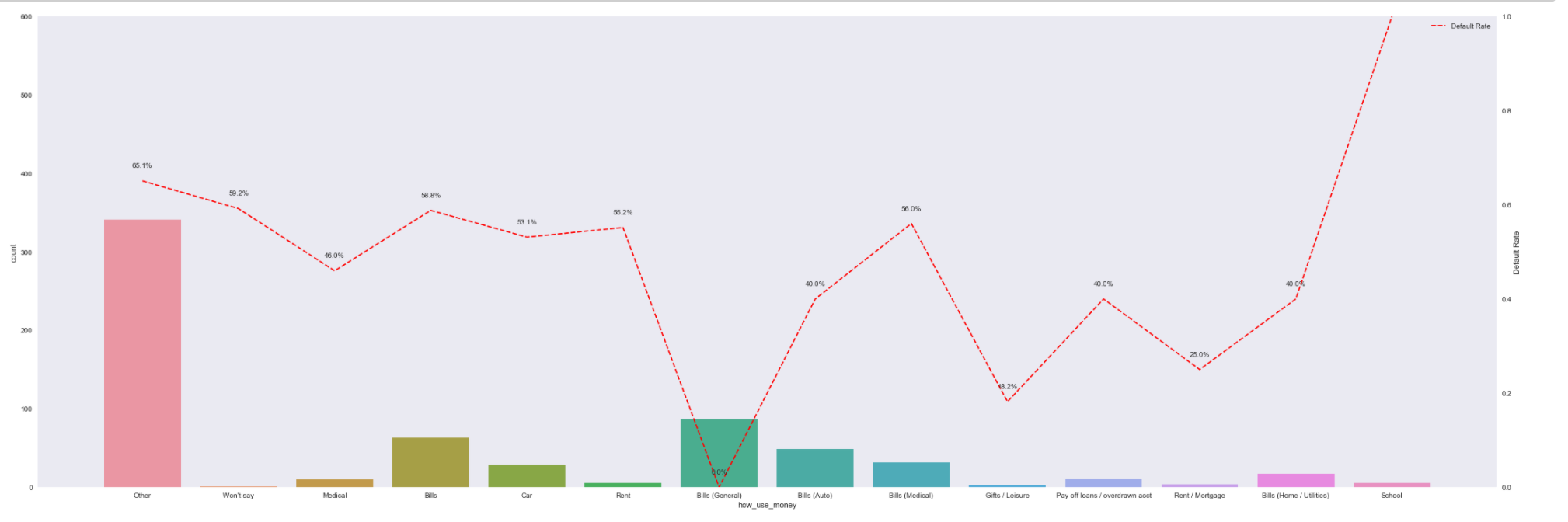
* Create a feature ‘Age’ indicating the age of the customer by “birth\_date“;
* Create a categorical feature ‘email’ indicating which email-service the customer uses (gmail, yahoo, hotmail etc.) by “email”
* Check how many missing values does each variable have
* Fill “payment\_amount\_approved” missing values by multiplying 90% of payment\_amount
* Create two Features: Monthly Income by Approved Payment Amount & Monthly Rent by Approved Payment Amount

1. Draw Plots to observe relationship between specific feature and default rate (high default rate means loan performance is bad)

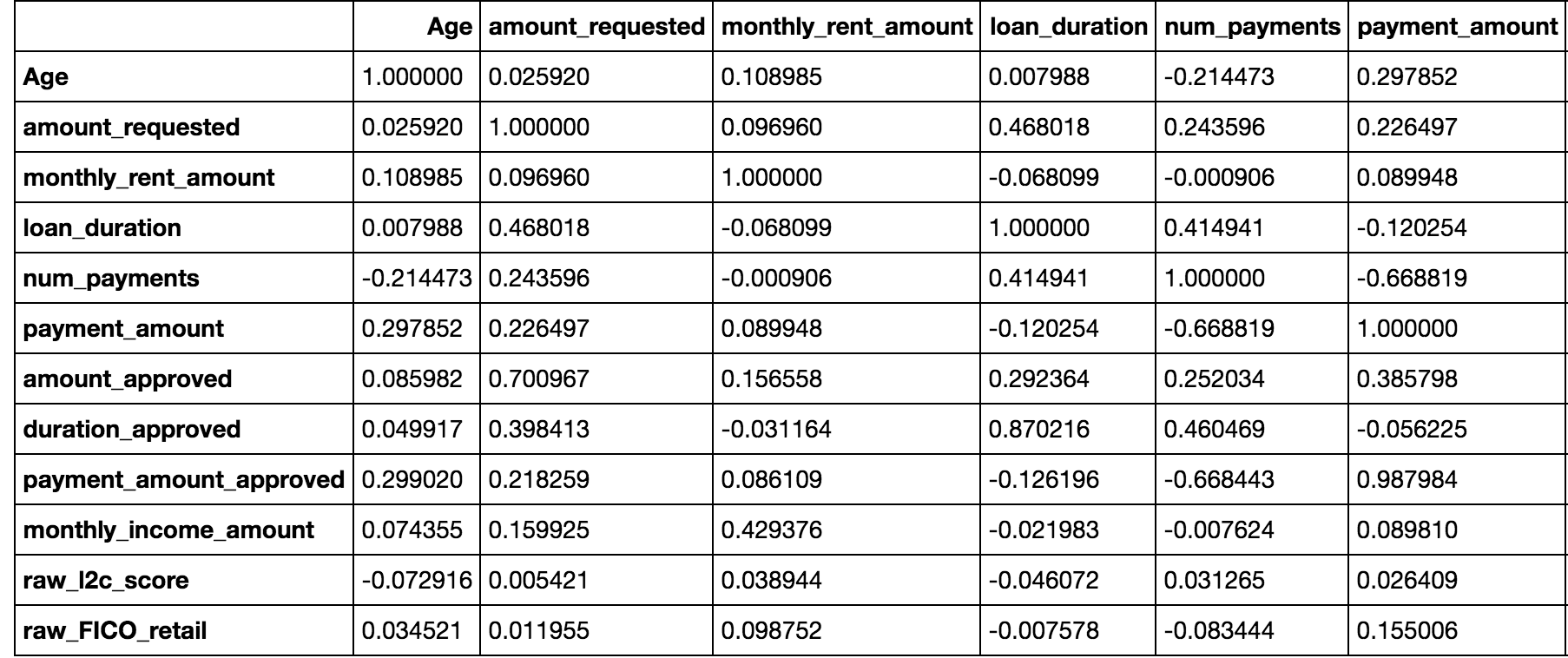
* Higher age, lower default rate
* Customer has low defaulte rate for those who save deposit directly into bank account
* Customer has low defaulte rate for those who don’t rent a house
* Customer has low defaulte rate for those who pay semi-monthly or monthly
* Customer has low defaulte rate for those who have high income
* Customer has lower defaulte rate for those who have high FICO score
* Customer has high defaulte rate for those who don’t specify how to use the loan







1. Calculate Correlation



1. Drop unnecessary variables

* Some variables are unnecessary in our analysis, like “bank\_routing\_number” and “status”. So, we drop them.
* Drop some varibles that have high correlation with others

1. Convert data types

* We need change our data types to the type that our machine learning algorithm can regonize it. So, we change the boolean type to int and categorical variables to dummy variables.

**Third Step – Try Different Machine Learning Algorithms in Scikit Learn**

11. Split the data into training set and test set into 50%:50%

* We use train data to fit the model and test data to check the accuracy.

12. Train the model and use feature selection to select variables

* Try different regularization parameter and see which gives the highest AUC score

13. Try different algorithms

* We want to try different machine learning algorithms and figure out which model gives the best accuracy. So, we try Logistic Regression, Gaussian Naïve Bayes, and Support Vector Machine with Gaussian kernel. The reason we want to use SVM with Gaussian kernel is that both number of features and number of observations are small.
* The result shows us that Logistic Regression model gives us 68.4% accuracy; Gaussian Naïve Bayes model gives us 67.9% accuracy; SVM with Gaussian kernel gives us 52.18% accuracy.
* As a conclusion, we want to use regularized Logistic Regression model to predict whether our new customers should be given a loan in the future.