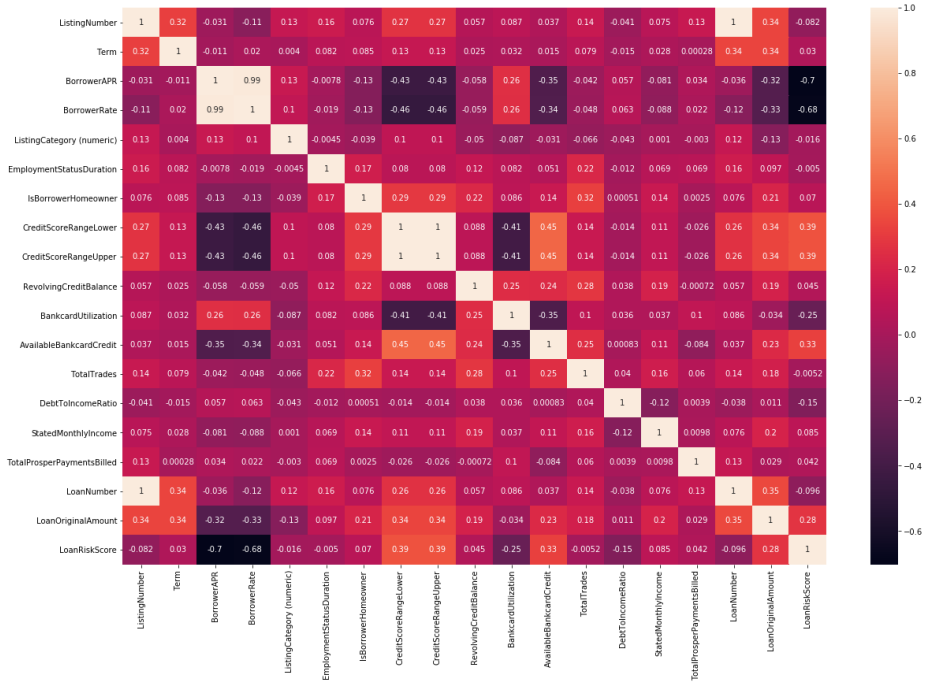
Milestone1

1) Preprocessing techniques:

- **Drop columns:** we remove two columns: (CreditGrade, TotalProsperPaymentsBilled) from the dataset because they have a lot of null values.
- **Replace Nonvalues:** we replace nonvalues from columns with mean of columns.
- **Encode the text features:** we convert some features from text to numerical values: (LoanStatus, BorrowerState, EmploymentStatus, IsBorrowerHomeowner, IncomeRange).
- **Feature Scaling:** we use standard scaler to scale some features: (RevolvingCreditBalance, AvailableBankcardCredit, StatedMonthlyIncome, LoanNumber).

2) Feature Selection:

- we get the features that affect LoanRiskScore by more than 30%
- The correlation matrix:



3) Regression techniques:

- **Model 1:**
  - Polynomial Regression
  - MSE testing: 1.6375833151045402 s
  - MSE training: 1.6339222023696616 s
  - Training time: 1.6735265254974365 s
  - Accuracy: 63.89266221570586 %
- **Model 2:**
  - Multivariable Regression
  - MSE testing: 2.225830068875934 s
  - MSE training: 2.2462563222099488 s
  - Training time: 0.045876264572143555 s
  - Accuracy: 63.378050758043344 %

4) The discarded features:

- **The messing features:** (ListingNumber) because it doesn't affect on dataset.
- **The dropped features:** (CreditGrade, TotalProsperPaymentsBilled) because they have a lot of nonvalues.

5) Splitting Dataset:

- We are split dataset to 70% training and 30% for testing.

6) Conclusion:

- We first clear data by many several preprocessing techniques to be more accurate to learn with different types of models.
- We note that polynomial regression takes training time more than linear model and take more time as we raise the degree.
- The MSE in linear model more than polynomial regression and accuracy is less.

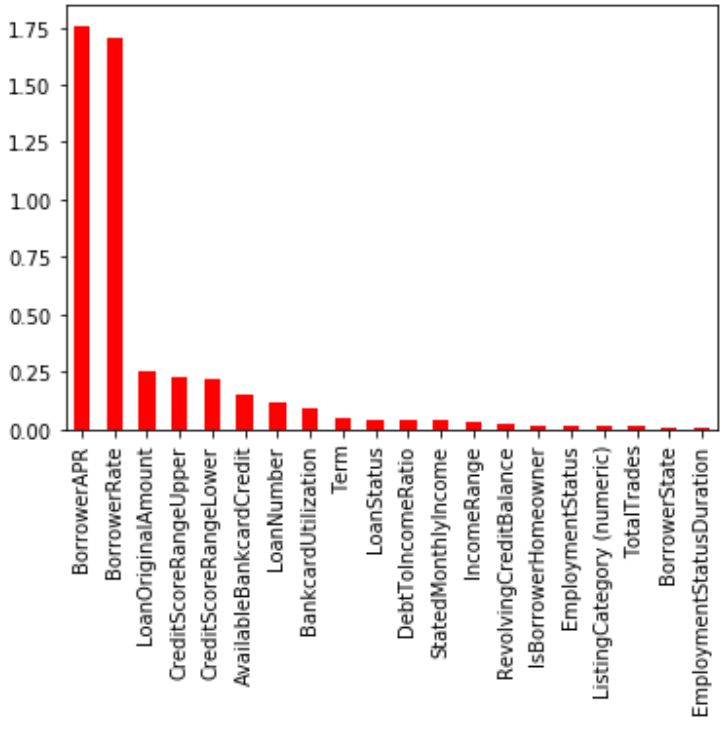
Milestone2

7) Preprocessing techniques:

- **Drop columns:** we remove two columns: (CreditGrade, TotalProsperPaymentsBilled) from the dataset because they have a lot of null values.
- **Replace Nonvalues:** we replace nonvalues from (ProsperRating (Alpha)) with (bfill) and other columns with (mean).
- **Encode the text features:** we convert some features from text to numerical values: (LoanStatus, BorrowerState, EmploymentStatus, IsBorrowerHomeowner, IncomeRange).
- **Feature Scaling:** we use standard scaler to scale some features: (RevolvingCreditBalance, AvailableBankcardCredit, StatedMonthlyIncome, LoanNumber).

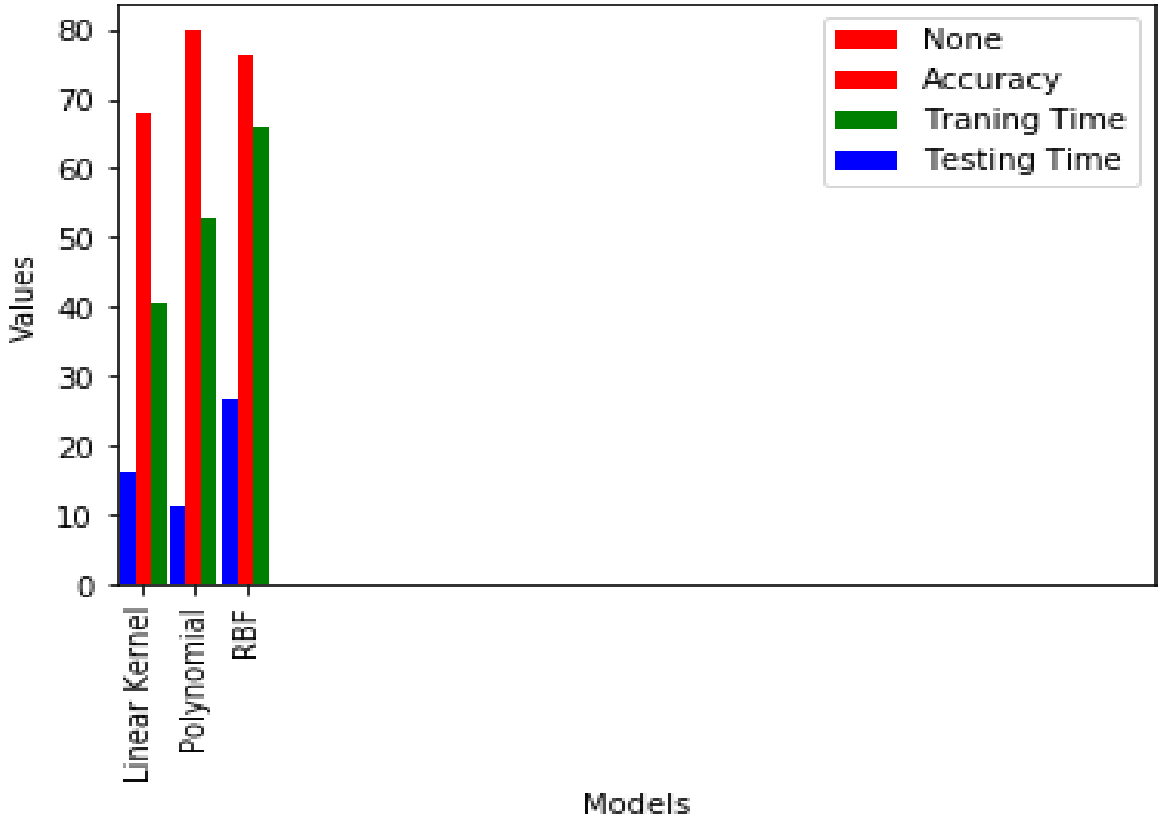
8) Feature Selection:

- We get the features that have more interconnection (BorrowerAPR, BorrowerRate).
- Info Plot:



9) Summarize Classification:

- **Model 1:**
  - Linear SVC Kernel
  - Accuracy: 67.94999378032094 %
  - Training Time: 40.68949031829834 s
  - Testing Time: 40.68949031829834 s
- **Model 2:**
  - polynomial SVC Kernel
  - Accuracy: 79.68652817514617 %
  - Training Time: 52.97307205200195 s
  - Testing Time: 11.471295595169067 s
- **Model 3:**
  - RBF SVC Kernel
  - Accuracy: 76.52693121034955 %
  - Training Time: 66.12568020820618 s
  - Testing Time: 26.553483486175537 s



10) Splitting Dataset:

- We are split dataset to 80% training and 20% for testing.

11) hyperparameter tuning:

- **Kernels:** After several times of training models, we note that:
  i. Accuracy: polynomial > RBF > Linear
  ii. Time training: RBF > polynomial > Linear
- **C:** We note that the less 'C' less than 0.1, accuracy be less, and training time be more. The more 'C' most than 1, accuracy still fixed, and training time be more.
- **Degree:** We note that the less 'Degree' less than 3, accuracy training time be less. The more 'Degree' most than 3, accuracy be most, training time be more, and testing be less (Over fitting).

12) Conclusion:

Classify a loan into one of seven categories: {A, AA, B, C, D, E, or HR} based (BorrowerAP and BorrowerRate) features with three Classification models Linear SVC Kernel, polynomial SVC Kernel, and RBF SVC Kernel. we found the best classification model from the three models is a polynomial model.