Predict Loan Interest Rate

This was part of a work assignment of a **Data Scientist** job application. In this assignment, I completed a mini data science project that involves end-to-end pipeline from data cleaning, data preparing, featuring exploration and engineering, model prototyping and selection, and evaluation. The goal of the project is to predict the interest rate of loan applications. Each loan application contains the following data:

X1	Interest Rate on the loan
X2	A unique id for the loan.
X3	A unique id assigned for the borrower.
X4	Loan amount requested
X5	Loan amount funded
X6	Investor-funded portion of loan
X7	Number of payments (36 or 60)
X8	Loan grade (A, B, C, D, E, F, G)
X9	Loan subgrade (A1, A2, A3, A4, A5, B1, ..., G5)
X10	Employer or job title (self-filled)
X11	Number of years employed (0 to 10; 10 = 10 or more)
X12	”Home ownership status: RENT, OWN, MORTGAGE, OTHER.”
X13	Annual income of borrower
X14	”Income verified, not verified, or income source was verified”
X15	Date loan was issued
X16	Reason for loan provided by borrower
X17	”Loan category, as provided by borrower”
X18	”Loan title, as provided by borrower”
X19	First 3 numbers of zip code
X20	State of borrower
X21	”A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income.”
X22	The number of 30+ days past-due incidences of delinquency in the borrower’s credit file for the past 2 years
X23	Date the borrower’s earliest reported credit line was opened
X24	Number of inquiries by creditors during the past 6 months.
X25	Number of months since the borrower’s last delinquency.
X26	Number of months since the last public record.
X27	Number of open credit lines in the borrower’s credit file.
X28	Number of derogatory public records
X29	Total credit revolving balance
X30	”Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.”
X31	The total number of credit lines currently in the borrower’s credit file
X32	”The initial listing status of the loan. Possible values are – W, F”

The input is a csv file with 400,000 lines where each line contains columns X1 through X32. X1 contains the Interest rate, and X2 through X32 contains data from which useful feature/variables can be extracted to predict X1.

1 Data Cleaning

The file `Data for Cleaning & Modeling.csv` contains one line of header followed by 400,000 lines of records. All 400,000 lines contains columns X1 through X32 except one line with only X1 (i.e. interest rates), and X2 through X32 are missing. This line was ignored.

A sample of a line from the file looks like this:

```
12.19%,54731,82364,"$15,000","$15,000","$9,080", 36 months,B,B4,Job title,< 1 year,RENT,85000,VERIFIED
- income,Aug-09,"Reason provided, by borrower",debt_consolidation,Debt consolidation for on-time
payer,12345,CA,12.13,0,Feb-00,0,,,10,0,28854,52.10%,42,f
```

- Although the file is declared to be in column-separated-value format, there is some subtle thing to be taken care of: the columns X4, X5, and X6 contains the amount of loan in the format `"$([0-9]+)(,[0-9]+)+"` (e.g. "$15,000"). The comma can mess up with the the delimiting comma between columns, and it needs to be converted to pure format (15000).
- Other than columns (X4, X5, X6), X10 (job title) and X16 (reason) may contain commas that can be mess up with the delimiting comma.
- Columns X10 (job title of borrower) and X16 (reason of loan) contains the unstructured text data, and some contains messing commas and are not easy to parse, and it's complicated to convert to structured data. For simplicity, X10 and X16 will be ignored in the current model.
- Columns X15 and X23 contains dates (month-year), should be converted to the number of months prior to Aug-2016.

After data cleaning, we end up with 338,851 out of 400,000 data records. The next step is to investigate which feature/variables are correlated with the target value.

2 Feature Exploration

As described earlier, each data record contains the target value (X1) along with features (X2 through X32). Apparently, not all features are necessary for training an adequate model for predicting X1. We need to identify which features are informative. First, the types of value (Numerical, Categorical, Ordinal) of each feature is summarized in the following table.

Type	Feature
Numerical (Continuous)	X4, X5, X6, X13, X21, X23, X29, X30
Numerical (Discrete)	X22, X24, X25, X26, X27, X28, X31
Categorical	X7, X12, X14, X32
Ordinal	X8, X9, X11
Ignored	X2, X3, X10, X15, X16, X17, X18, X19, X20

2.1 Features that are ignored

- X2 and X3 contains nothing but ID numbers for loan applications and borrowers. It seems that no repeated IDs were found (i.e. no borrower has more than one loan applications). It's not likely that IDs can be correlated with interest rates.
- X19 and X20 contains geographical and demographic information. To investigated if they are correlated with interest rates, the records in the training date are sorted with respect to interest rates. Consider the top and bottom 10% of records (i.e. High and Low interest rates). Find the set of zip code that occur most frequently (at least 100 times) in the two groups. We end up with 91 zip code in the high set and 104 in the low set. However, the overlap between them is observed to be 84, which suggests that zip code does not provide very informative/discriminative information.
- Columns X10, X16, X17, and X18 contains textual data. Some contain missing values because they are optionally filled by the borrower. For simplicity, they are ignored in the downstream analysis.

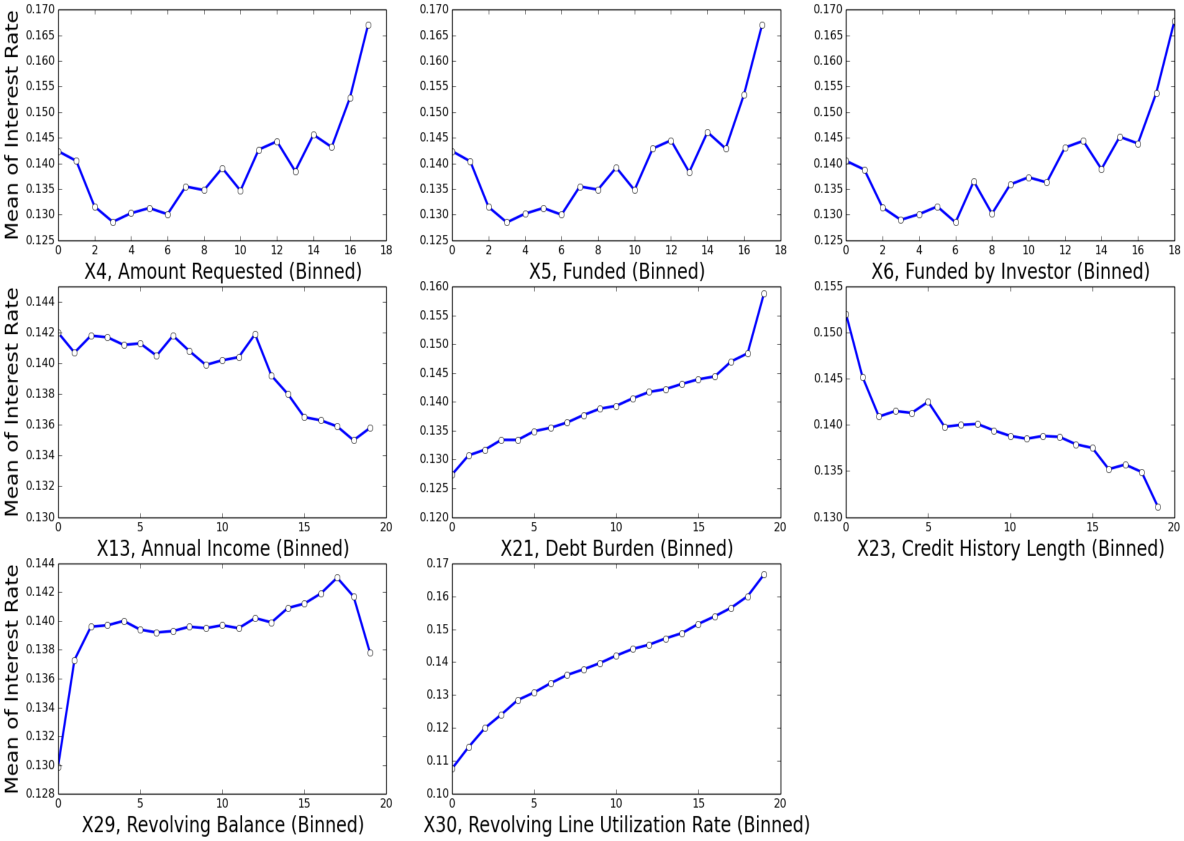

Figure 1: **Interest rate vs. numeric (continuous) features**

2.2 Features that are relevant

2.2.1 Numeric (continuous) Features

See Figure 1.

- X4, X5, X6: These columns contains information related to the amount of loan (request, funded, portion funded by investors). As we can see, interest rate initially drops as the loan amount increases, but shoots up quickly as the loan amount continuous to increase.
- X13: Interest rate decreases as annul income of borrow increases.
- X21: Interest rate increases as the burden of debt payment increases.
- X23: Interest rate decreases as the number of months of credit history increases.
- X29: Interest rate is low when the revolving balance (i.e. unpaid amount) is close to zero, but shoots up high and stays high as the revolving balance increases.
- X30: Interest rate consistently increases as the revolving line utilization rate (i.e. percentage of available credit used) increases.

2.2.2 Numeric (discrete) Features

See Figure 2.

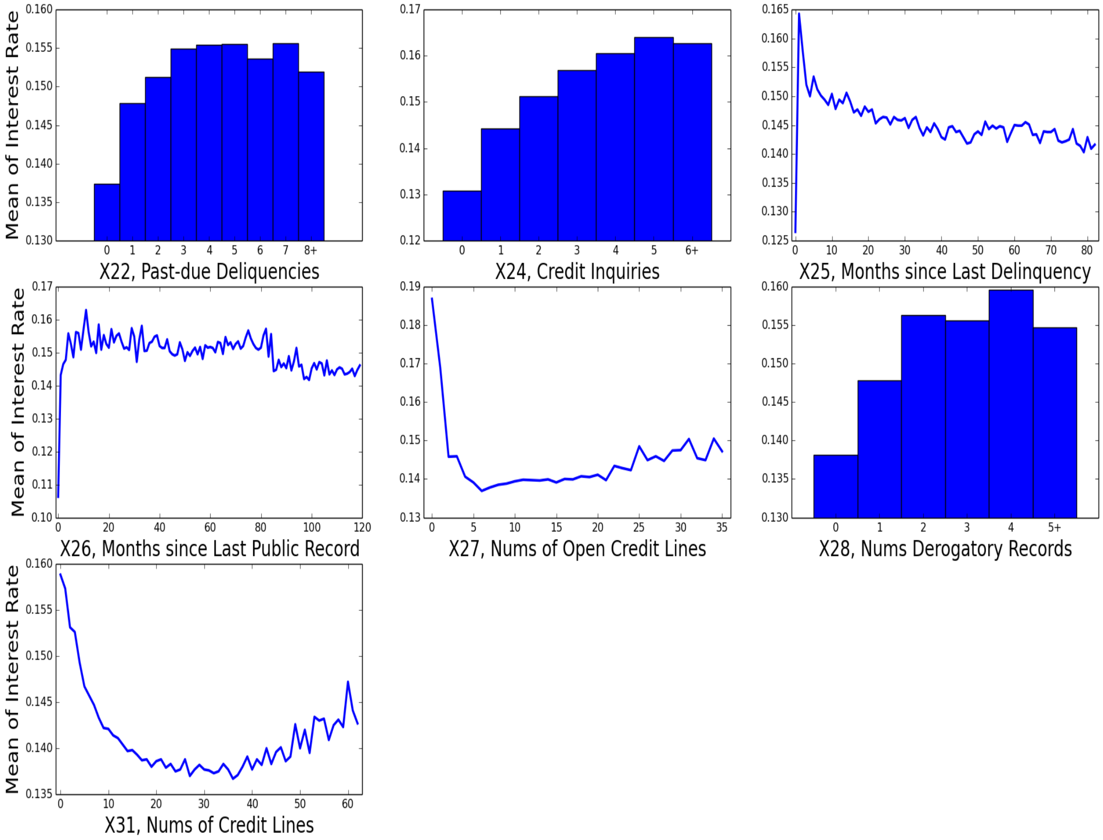

Figure 2: **Interest rate vs. numeric (discrete) features**

- X22: Interest rate is low when the number of past-due delinquencies is zero, and increases until the number of delinquencies becomes 3, and stays relatively the same onwards.
- X24: Interest rate increases consistently with the number of credit inquiries, and stays relatively unchanged when the number reaches 6.
- X25: Interest rate is very low when the borrower has a "clean sheet" (month since last delinquency equals 0), when shoots up very high when the delinquency incidence just occurred last month, and decreases as the delinquency incidence was less recent.
- X26: This looks like an "On-or-off" variable: interest rate is low when the borrower has no public record, but becomes consistently high at similar level when the borrow has public record.
- X27, X31: Interest rate is high when the borrower has no credit lines before, but decreases very quickly when the borrower has one or more credit lines. But when the borrower has too many credit lines, the interest rates starts growing.
- X28: Interest rate is low when the borrower has no derogatory records, but increases with the number of derogatory records, and stays relatively unchanged when the number becomes 3 or greater.

2.2.3 Ordinal Features

See Figure 3

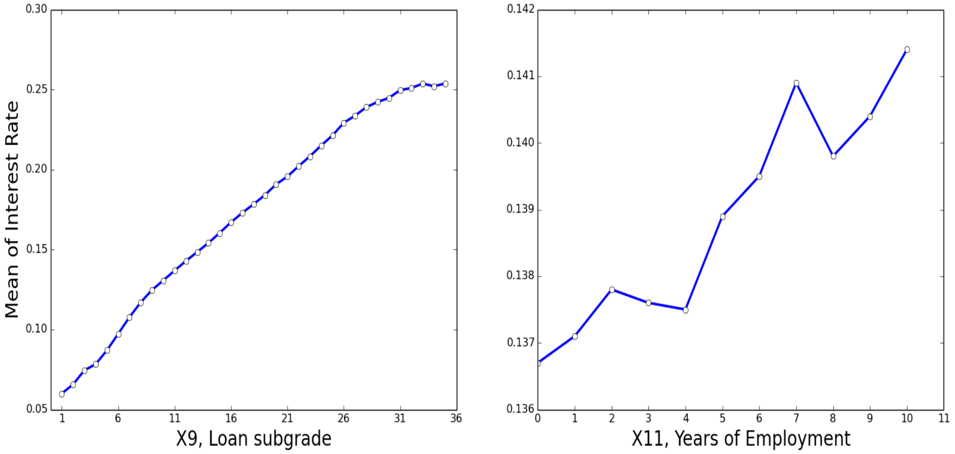

Figure 3: **Interest rate vs. ordinal features**

- X8, X9: Loan grade is denoted by A, B, C, D, E, F, G. Each grade is further divided into five sub-grades 1, 2, 3, 4, 5. We can used integers 1 through 35 to denote sub-grades A1 through G5. As we can see, the mean of interest rate grows approximately linearly with the corresponding integer of the loan sub-grade.
- X11: Interestingly, the interest rate grows with the number of years of employment of the borrower.

2.2.4 Categorical Features

See Figure 4.

- X7: The borrower would be given a higher interest rate if they chose the 60 month payment option
- X12: It seems that renters and owners have very similar levels of interest rate, except that there is trend: renters have slightly higher rate than owners without mortgage, followed by owners with mortgage.
- X14: This variable appears to be interesting: borrowers whose income or income source was verified tend to have higher interest rate than those whose income was not verified.
- X32: These two samples of interest rate (initial listing status = 'w' or 'f') appears to be almost the same. However, a two-sample t-test shows that the two samples of interest rates have statistically significantly different means (p-value = $1.138E-8$).

2.2.5 Missing Value Imputation

The problem of missing values is very common in real world dataset. Table 1 summarizes the number of missing values for each variable.

Missing values can be imputed in a number of ways. The most common approach is to replace the missing value with global mean/mode of the variable/feature that contains missing values. But in this dataset, we can take advantage of the correlation between features and the target value: first, sort the loan applications in ascending or descending order of the interest rate, then divide the sorted dataset into n bins, and the missing values in each bin can be computed as the the mean or mode of that variable in the corresponding bin (i.e. **conditional mean/mode**). However, if the number of non-empty values in each bins is too small, the conditional mean or mode may not be sufficiently representative. In this case, the missing values are imputed as the global mean or mode.

Figure 5 illustrates the means of the interest rate of each bin of the continuous value (X13, X22) or discrete value (X25, X26, X9, X11, X12) for the original data (blue) and the original plus the imputed data (red). As we can see, the imputed values appears to have similar distribution with respect to the original data.

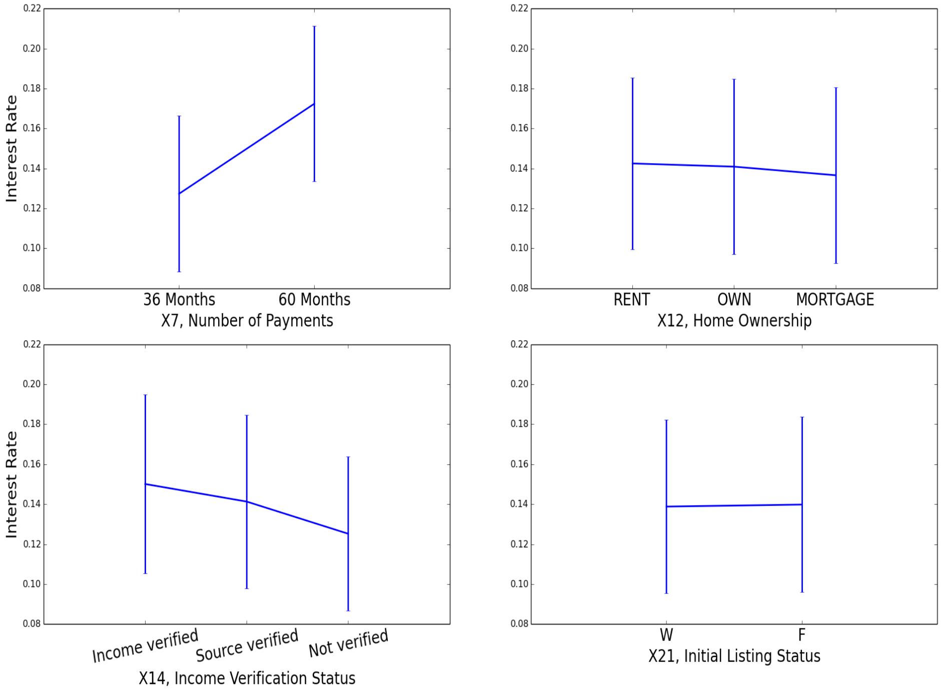

Figure 4: **Interest rate vs. categorical features**

Feature	Number of Missing Values
X9	51,850
X11	14,792
X12	51,959
X13	51,733
X25	185,382
X26	295,455
X30	222

Table 1: Summary of Missing Values

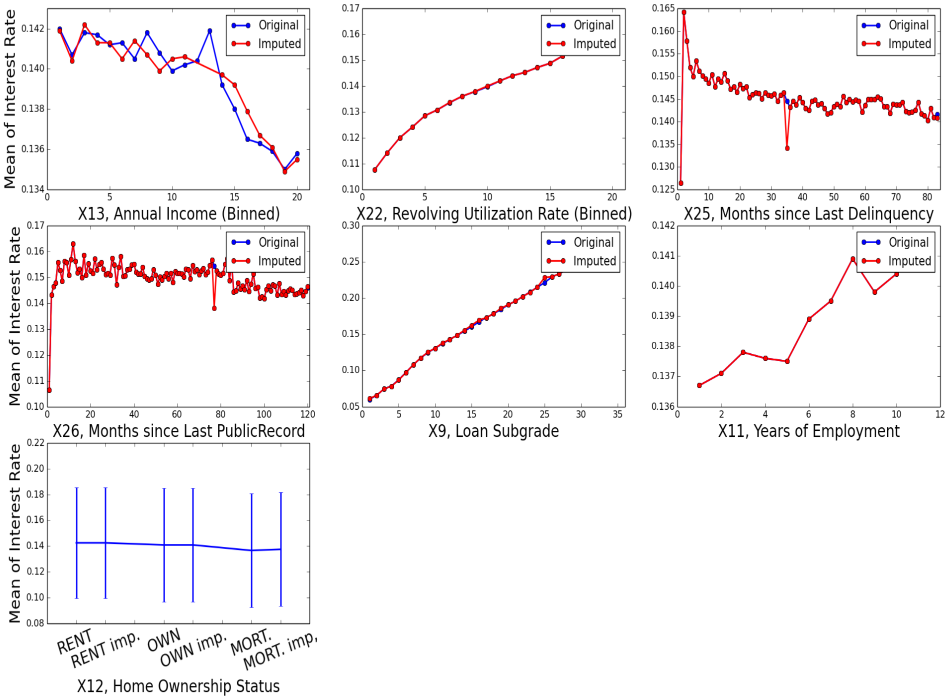

Figure 5: **Original vs. Imputed values**

2.2.6 Feature Representation

Most numeric values can be represented as is and do not need transformation. The features that need to be transformed include:

- X23: As mentioned earlier, features in *datetime* type should be converted into the number of months prior to Aug-2016.
- X25: Because the target value does not increase/decrease monotonically with X25, X25 is transformed as follows: $X_{25} = 0$, if $X_{25} = 0$ else $1/X_{25}$. This function mimics the distribution of interest rate with respect to X25 (Figure 1).
- Categorical features (X7, X12, X14, X32) are transformed to "one-hot" encoding scheme. For example, $X_7 = [1, 0]$ if X_7 equals "36", or $[0, 1]$ if X_7 equals "60".
- Discrete numeric features (X22, X24, X25, X26, X27, X31) often contains "counts" that have dramatically difference ranges. The empirical results (Figure 2) shows that the target value (interest rate) changes with X with a predictable pattern when X is small. Thus these variables are transformed as follows: $X = X$ if $X < n$, or n if $X \geq n$.

3 Results

The final cleaned and transformed dataset is represented as real-valued matrix **X** of 338851 rows and 27 columns. The 27 columns contains the continuous numeric features (X4, X5, X6, X13, X21, X23, X29, X30), discrete numeric features (X22, X24, X25, X26, X27, X28, X31), ordinal features (X9, X11), and categorical features (X7, X12, X14, X32). The categorical features are transformed into one-hot encoding scheme (X7:2, X12:3, X14:3, X32:2).

Random Forest Regression was used to predict interest rate. Ten-fold cross-validation was used to evaluate the performance, in which the set of examples were split into ten disjoint subsets, and the model trained on the union of nine subsets was evaluated on the remaining one subset.

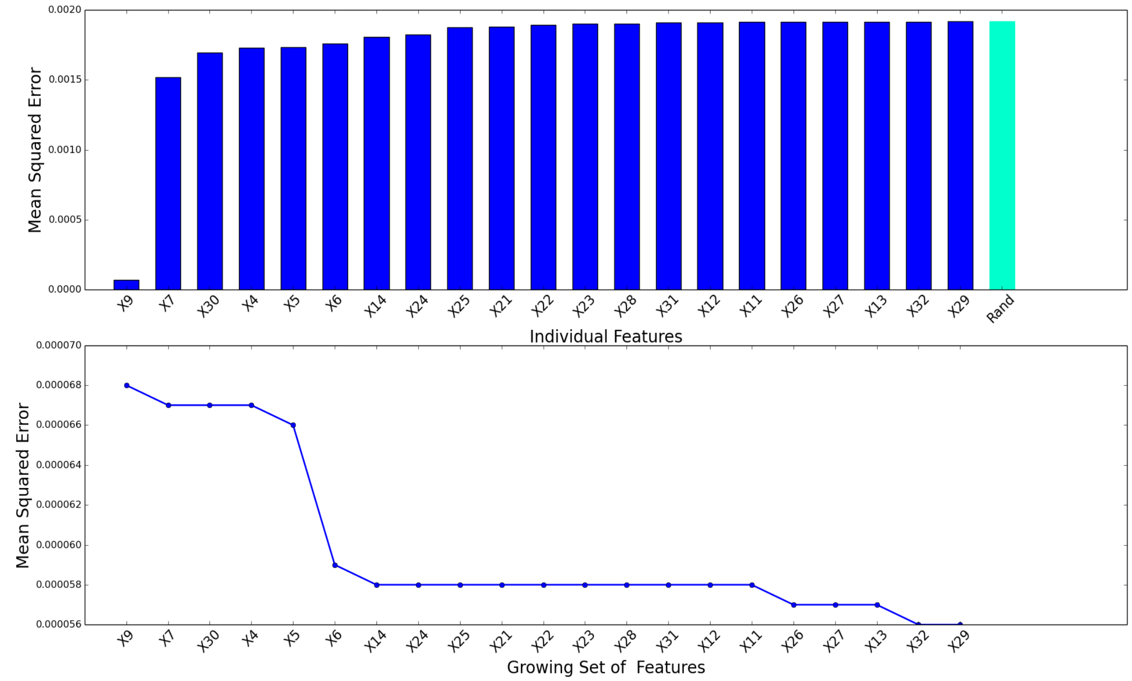

Figure 6: **Predictive power of features**

To investigate the predictive power of each individual feature, $\mathbf{X}$ was projected to each feature $\mathbf{x}_i$ on which a different model was trained. The MSE (Mean squared error) was computed on a vector of target values and a vector of predicted values.

All features were ranked in ascending order of their MSE. As shown in Figure 6, all features provide comparable or better discriminative power than the vector containing only random numbers (i.e. Cyan). In particular, the integer representation of the loan subgrade (X9) resulted in much better performance than all the other features.

Starting with only the feature X9, the other features were progressively added to X9 in ascending order of their MSE, and the model was evaluated on the growing set of features. Although X9 alone provides outstanding discriminative power, the model performance was improved when the less discriminative features were added.