* **Data clean and visualization**
* **using decision tree and random forest fit models**
* **using K mean method to cluster all data, figure out whether 7 or 8 category is better.**

**1. Data clean and visualization**

**- chose variables that matters.** I only choose what I believe are very important and independent. for example, since all of BorrowerAPR, EstimatedEffectiveYield, EstimatedLoss, EstimatedReturn are related or comes from the BorrowerRate, so I only select BorrowerRate.

**- find variables missing a lot of values** Then find out columns that has miss value and try to fill nan. For numerical variables, I use the normal distributed values from their exist value to fill the nan. For character variables, I will either assign them ordered number according their meanings, or I will combine some categories, or delete unnecessary data.

A picture containing drawing

Description automatically generated**- visualization**A screenshot of a cell phone

Description automatically generatedBefore I evaluate the credit grade, I want to have a look at these variables by plotting. (the target variable called credit grade before 07-01-2009, after then it called Prosper Rating (Alpha), they are different columns, but indicate same thing.)

The next picture is the count of borrowers` state, it`s easy to see, loan status will not have relationship with borrowers` state. So does the employment status.

Let`s look at other numerical value that might have relationship with loan status.

A screenshot of a cell phone

Description automatically generatedFrom the heatmap, credit score has negative relationship with bank card utilization, as well as income range, loan original amount, this is reasonable, since people have more money are more likely to pay all the bills.

**2. fit model**

I want to know: Whether new credit grade is a better indicator of important numerical features than old one: After using decision tree and random forest respectively, it is hard to say the new credit grade is more fit the numerical features.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | TP | FP | TN | FN | accuracy | TPR | TNR |
| tree\_before | 1605 | 6 | 1674 | 16 | 0.9852 | 0.9901 | 0.9964 |
| tree\_after | 1628 | 105 | 774 | 109 | 0.8389 | 0.9372 | 0.8805 |
| forest\_before | 1586 | 31 | 1650 | 23 | 0.9749 | 0.9857 | 0.9816 |
| forest\_after | 1602 | 128 | 758 | 151 | 0.7984 | 0.9139 | 0.8555 |

**3. build new credit grade to find out whether 7 grade is better or 8 grade**

using k-mean method, trying to build a new credit grade for data both before and after 07-01-2009.

Since there are two credit grade, one has 8 scales before 07-01-2009, the other has 7 scales after 07-01-2009, I`m trying to combine them and build a new scale for credit evaluation.

Using Elbow Method, I find k = 7, and then the result has a accuracy for k mean model is 0.6725, when k=8, the accuracy is 0.6724.

Conclusion, the older credit grade is better than new one at least from the variables I have choose. And whether to cluster all data by 7 is better than 8.