LOAN CREDIT RISK PREDICTION

Author: Rania Virda Sukmaningsih

Background

Loan companies are faced with two major decisions that carry **two types of risk** regarding approval decisions:
1. **Approving loans** to applicants who are unlikely to **repay their loans** resulting in **financial losses** for the company.
2. **Disapproving loans** to applicants who are likely to **repay the loan** resulting in **business losses**.

To reduce this credit risk, it is necessary to assess whether **the applicant** is a **good loaner** or a **bad loaner**.

Dataset & Bussiness Understanding

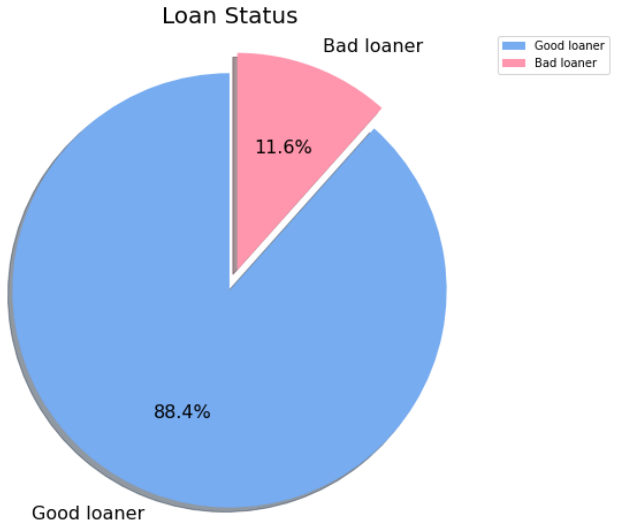
Dataset information:
This dataset contains information on loan lending from a lending company, namely LendingClub from 2007 to 2014.

Attribute Information:

- **Identifier:**
  `id` and `member_id` is unique LC ID that each of which is an ID for loan listing and ID for the loaner member
- **Target:**
  `loan_status` has several values, such as:
  - `Current` means current payments
  - `Charged Off` means the payment is in default so that it is written off
  - `Late` means late payment is made
  - `In Grace Period` means in grace period
  - `Fully Paid` means payment in full
  - `Default` means payment is stuck

  Later loan_status will be categorized as good loaner and bad loaner.
- **Company Goals:**
  - Accepting applicants who will be good loaner
  - Declining applicants who will be a risky borrower or bad loaner
- **Problems:**
  Loan companies are faced with two major decisions that carry two types of risk regarding approval decisions:
  - Approving loans to applicants who are unlikely to repay their loans resulting in financial losses for the company
  - Disapproving loans to applicants who are likely to repay the loan resulting in business losses
- **Objectives:**
  - Predict whether the applicant is a good loaner or a bad loaner
  - Whats makes the borrower indicated a bad loaner

What Happened?

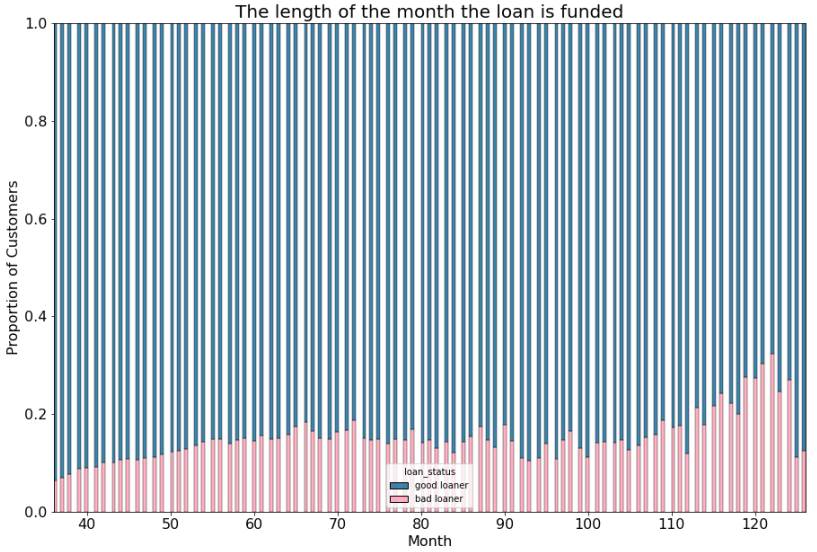


Good loaners is when the loan status is **current, fully paid, late < 30 days, & does not meet the credit policy with status fully paid.** Otherwise is **Bad loaners** (such as charged off, in grace period, late > 30 days, does not meet the credit policy with status charged off).

Who is applying for a loan?

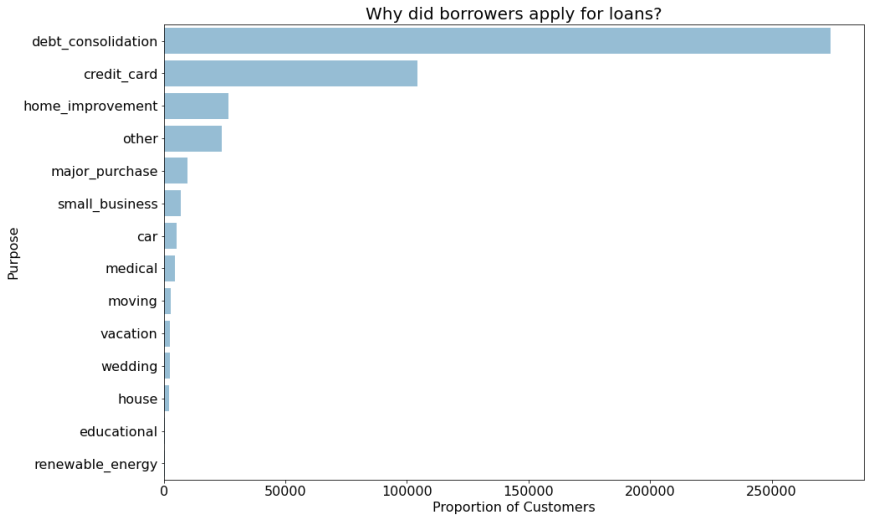
- In employment title, most of applicants have **job** in **Manager, Service, Asistant**.
- Many aplicant didn't write their employment title, so it's marked **nan**

How long the loan is funded?



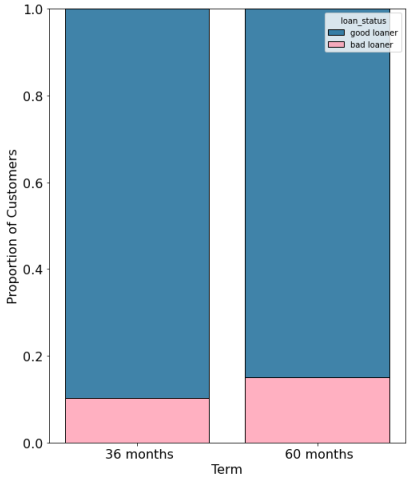
123 months before December 2017, the highest percentage of **bad loaners** is indicated.

Why is applying for loan?



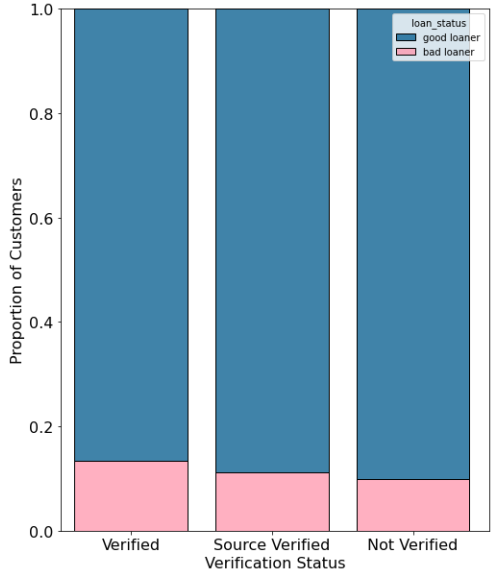
Most of borrower's applied loan **for dept consolidation**

Loan status by term



Loan risk with a span of **60 months** has a **high level of adverse impact**

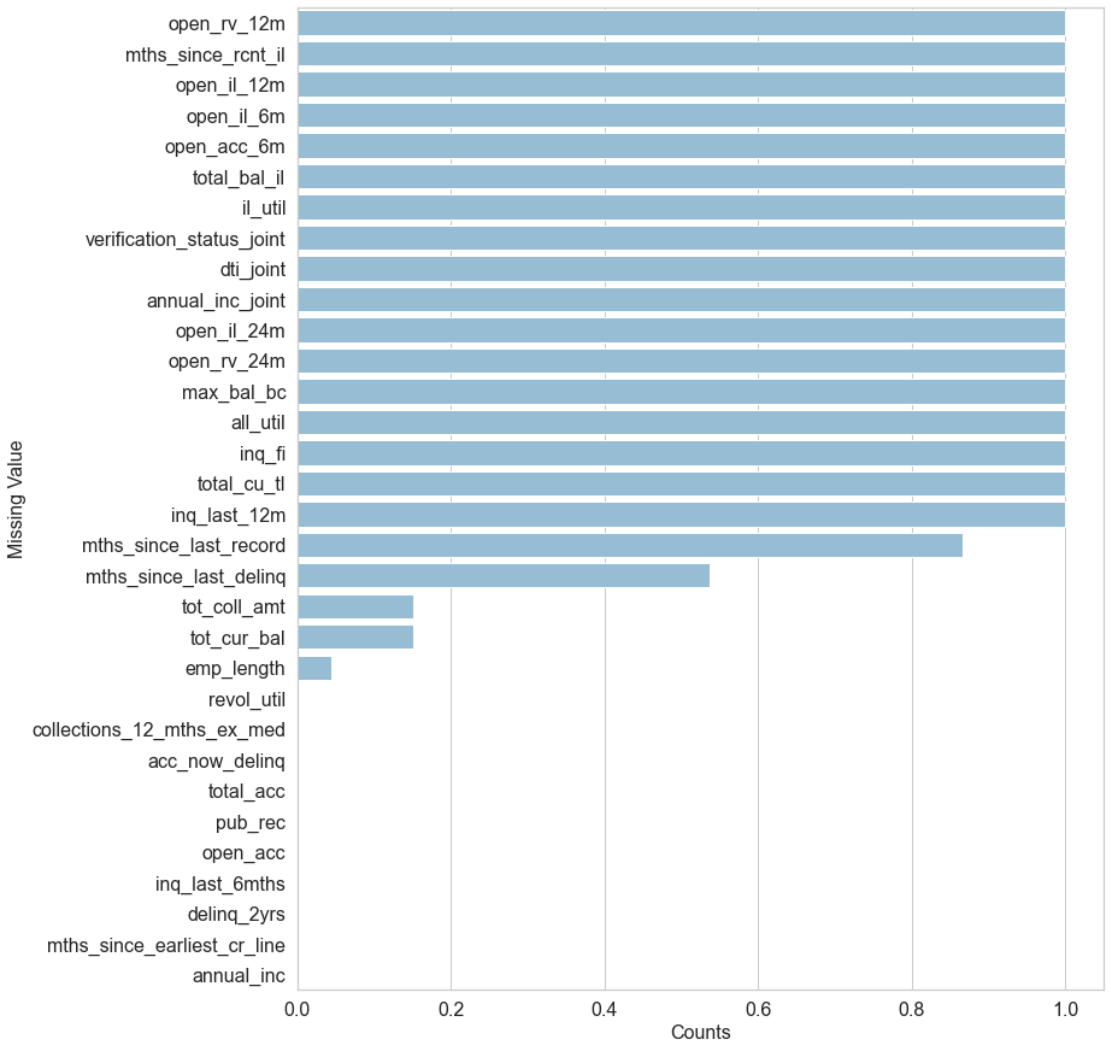
How many have been verified?



Bad loaners indicated the most on **verified status**.

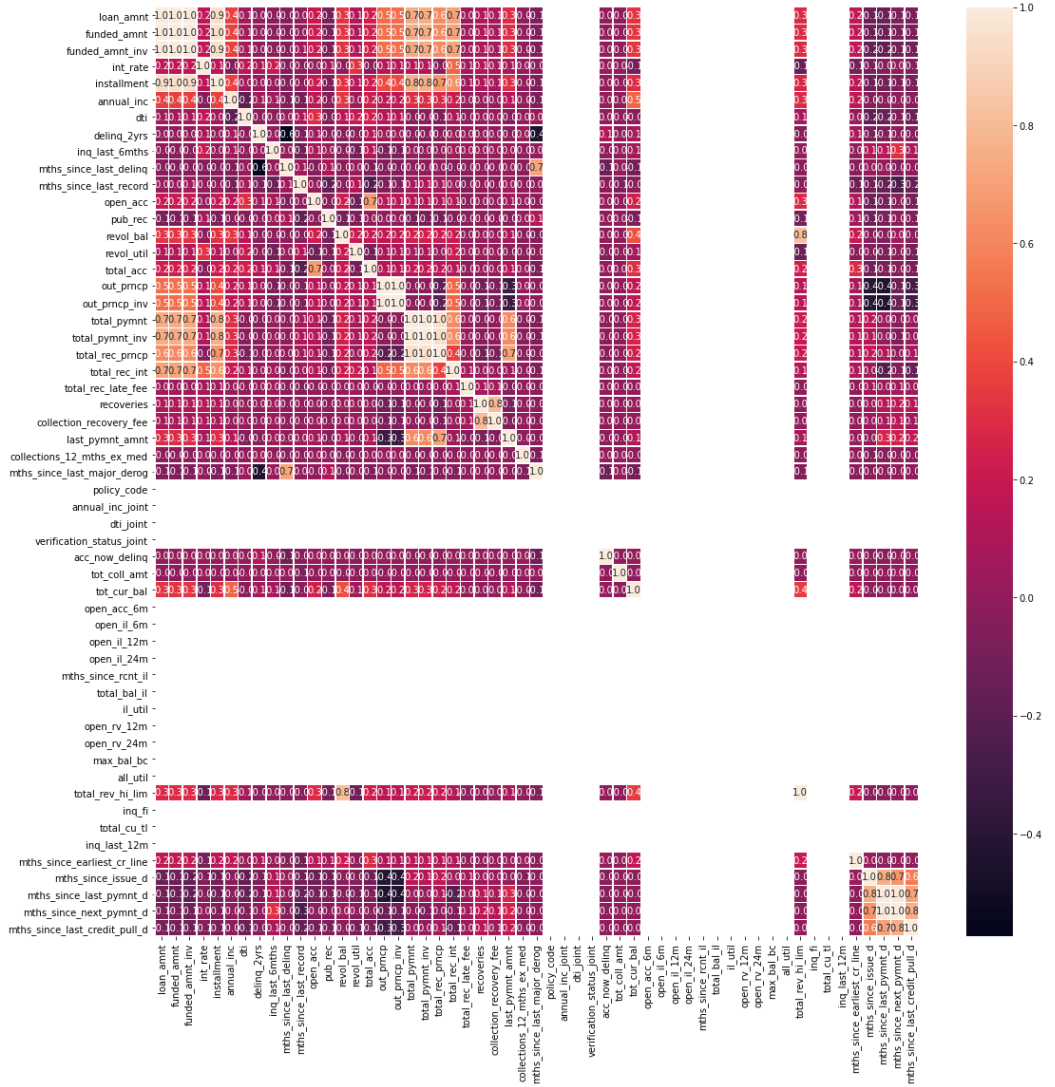
id/x partners By: Rania Virda Sukmaningsih

Missing Values



There are lots of features that contain null values, if those features contain **>75% features**, those features **will be deleted**. However, if the null values on the features are **< 75%,** it will be **imputed with mode or mean.**

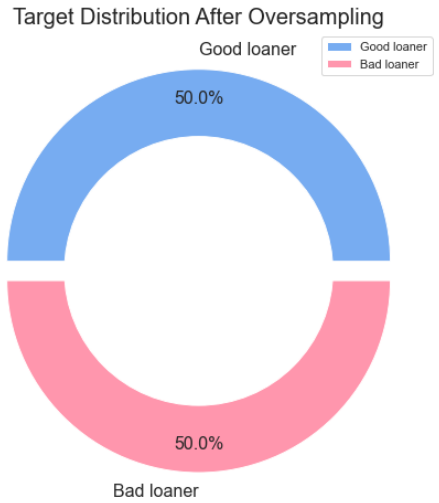
Feature Selection



Here, if there are pairs of features that **have a high correlation**, only **one will be taken**. The correlation value that is used as a benchmark as a high correlation is uncertain, generally the number **0.7** is used. In addition, there are lots of features that all intend to have a **null value**, so they **will be removed.**

id/x partners By: Rania Virda Sukmaningsih

Oversampling with SMOTE



This dataset is imbalanced. I use SMOTE to make it balanced.

Model Development

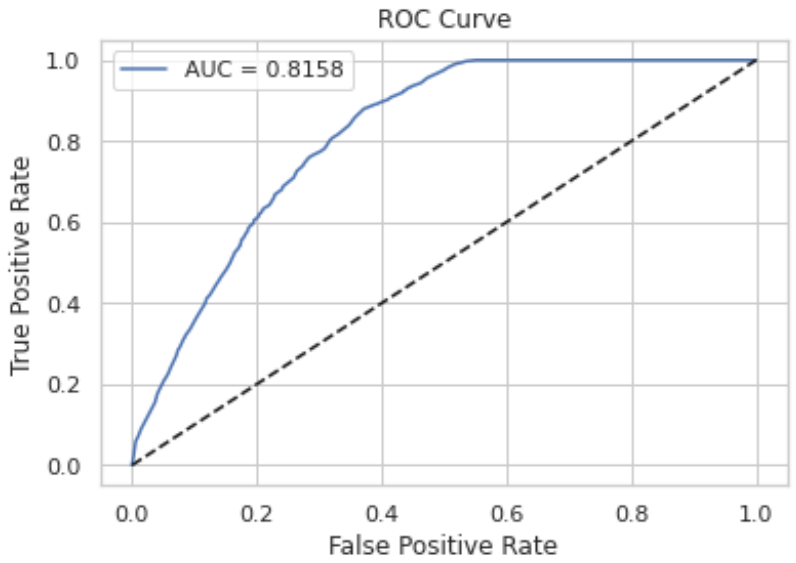
I use Logistic Regression, Random Forest. Naive Bayes, Perceptron, Stochastic Gradient Decent, Linear SVC, and Decision Tree for model development.

Model Evaluation

- I want to **avoid** either high **false negatives or high false positives**, therefore I will use the **F1 score** for model evaluation
- I'm still **paying attention** to the **accuracy score** as well since this metric is easier to interpret
- I'm also using **cross validation performance** to **estimated accuracy score** for data validation with 3-folds.
- In credit risk modeling, **test performance** is calculated using the **AUC metrics**.

Model	Accuracy	F1 Score	Cross Validation Score (3-folds)
Random Forest	80.36	0.834650	0.936101
Logistic Regression	76.74	0.883966	0.936069
Linear SVC	76.64	0.899879	0.925014
Decision Tree	76.49	0.267605	0.936562
Stochastic Gradient Decent	75.24	0.920789	0.919566
Naive Bayes	74.24	0.889939	0.930311
Perceptron	67.44	0.829503	0.873361

Random forest classifier give the highest performance



ROC Curve performance reach **0.8158** using random forest classifier

Conclusion

- **Best model: Random Forest** with 3-folds cross validation
- The **test** was carried out using the **AUC metric** with a **random forest model.** The resulting **AUC score** is **0.82**, which includes **good performance** in credit risk modeling.
- We should **pay more attention** to borrowers who meet the **criteria :**
  - **earlier issue date**
  - **loan application within 36 months**

Click here to see my code

id/x partners By: Rania Virda Sukmaningsih