Insights into LendingClub.com Client Loan Portfolio:   
An Analytical Breakdown

Key words: Lending, Borrowing, Supervised Learning, Unsupervised Learning, Statistics, Clustering

1. Introduction

The project analyses the data provided by [LendingClub.com](https://www.linkedin.com/company/lending-club/), a financial services company.   
It is divided into two parts:

1. **Unsupervised Learning**: this involves clustering clients based on their characteristics when they first approach the company. The goal is to identify specific groups of clients who are more likely to receive a loan.
2. **Supervised Learning**: this part aims to understand the variables that have the greatest influence on the investors and on the company's decision-making process regarding the client, such as the interest rate on the loan and whether the client complies with the company's credit policy. Subsequently, predictive models are created to forecast the company's decisions, with the best model(s) being evaluated for effectiveness.

The data was processed using R programming language within the Rsudio environment. Version control was conducted through [GitHub](https://github.com/chiesastefano/loanData?tab=readme-ov-file).

2. Data Used

2.1 Source and Columns Outline

The [dataset](https://www.kaggle.com/datasets/saramah/loan-data?resource=download), uploaded by [Sara Mahdavi](https://www.kaggle.com/saramah) on Kaggle in CSV format, contains lending data from 2007 to 2010. Here is an outline of the columns:

**Categorical Variables:**

* credit.policy: binary variable indicating whether the customer meets the credit underwriting criteria of LendingClub.com (1 for meeting criteria, 0 otherwise);
* purpose: the purpose of the loan, with categories including *"credit\_card*", "*debt\_consolidation*", "*educational*", "*major\_purchase*", "*small\_business*", and "*all\_other*";
* not.fully.paid: Binary variable indicating whether the loan is not fully paid (1) or is fully paid (0).

**Numerical Variables:**

* int.rate: the interest rate of the loan;
* log.annual.inc: the natural log of the borrower self-reported annual income;
* installment: the monthly installments owed by the borrower if the loan is funded;
* dti: the debt-to-income ratio of the borrower, calculated as the amount of debt divided by annual income;
* fico: the FICO credit score of the borrower;
* days.with.cr.line: the number of days the borrower has had a credit line;
* revol.bal: the borrower's revolving balance, i.e., the amount unpaid at the end of the credit card billing cycle;
* revol.util:this represents the borrower's revolving line utilization rate, which indicates the proportion of the credit line used relative to the total credit available, expressed as a percentage. Generally, a higher utilization rate suggests that the client relies more heavily on debt, posing greater risk;
* risk.inq.last.6mths: the borrower's number of inquiries by creditors in the last 6 months.
* delinq.2yrs: The number of times the borrower has been 30+ days past due on a payment in the past 2 years;
* pub.rec: the borrower's number of derogatory public records, such as bankruptcy filings, tax liens, or judgments.

The dataset comes with 9488 records, with each line corresponding to a user of the website.

2.2 Data Manipulation

Utilising the column provided by the dataset three other columns have been created:

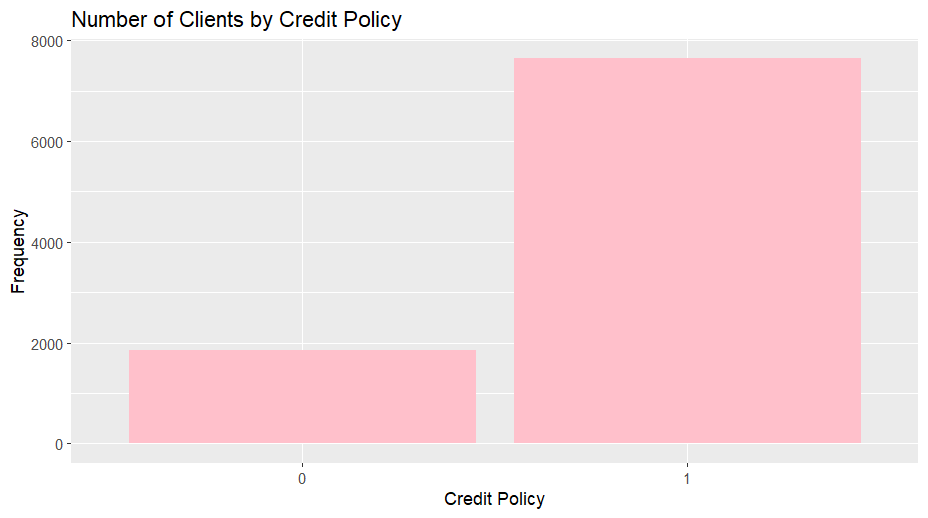
* annual.inc: exponential of the natural logarithm of the annual income (log.annual.inc);
* debt: product of the annual income (annual.inc) and the debt-to-income ratio (dti);
* total\_interests: product of the debt and the interest rate;
* iti: installment to monthly income ratio.

3. Data Visualization

Title: Exploring Data Through Visualization

The objective is to delve into the data, seeking out useful features for answering to our questions.

Borrowers will be referred to as “Clients” sometimes.



A high majority of the borrowers (80.5%) respect the underwriting criteria of LendingClub.com, while the others (19.5%) don’t. This variable will be used to compare different clients, to understand if there’s a difference between clients that respect the policy and the others.

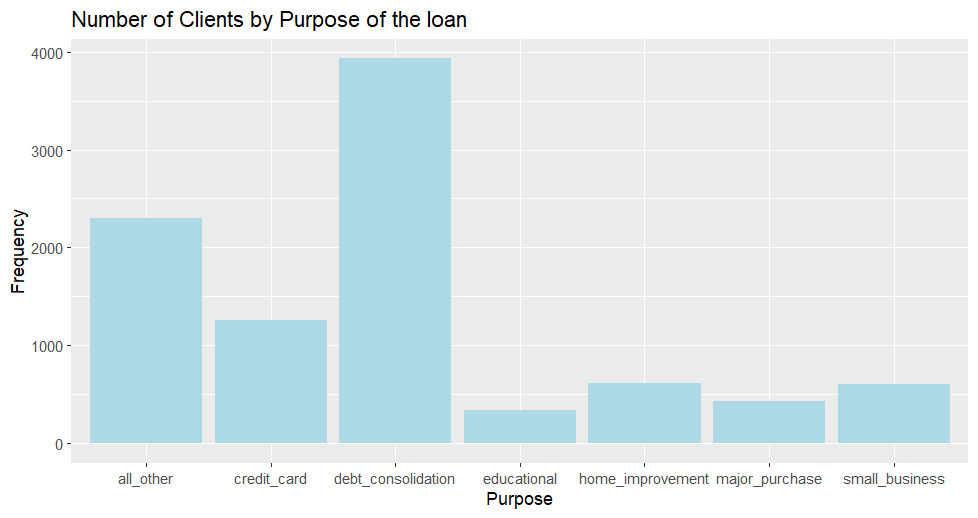
Most of the users ask for a loan for debit consolidation, followed by credit card payments, home renovations, small business funding, major purchases, educational purposes and others.

Immagine che contiene schermata, testo, diagramma, Diagramma

Descrizione generata automaticamente

The interest rate data looks like a right-skewed distribution. The median interest rate is 0.1221. The 75% of the records presents an interest rate lower than 0.14.

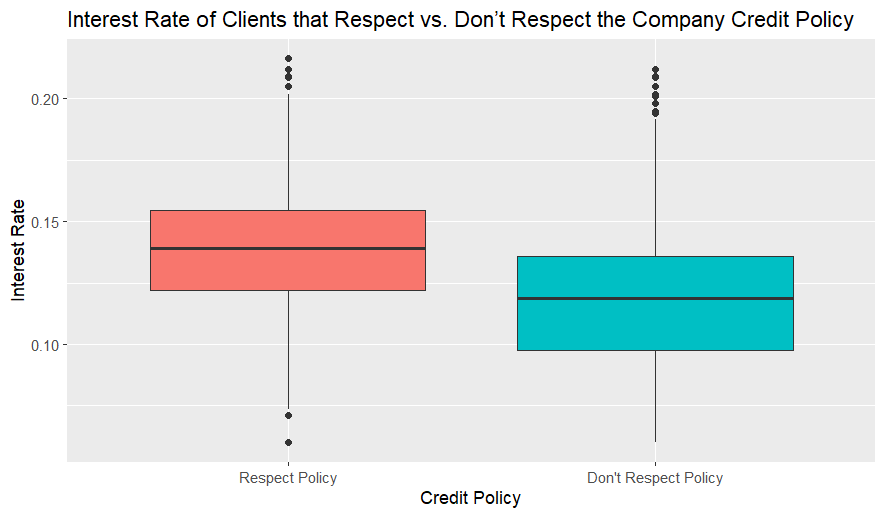
The Interest rate is higher for the clients that respect the company policy.

Immagine che contiene schermata, diagramma, testo, Diagramma

Descrizione generata automaticamente

The Installment, the monthtly payment if the lending is funded, has a right-skewed distribution. The median is 268.42 and the 75% of the records have an installment lower than 428.65.

Immagine che contiene schermata, testo, diagramma, Diagramma

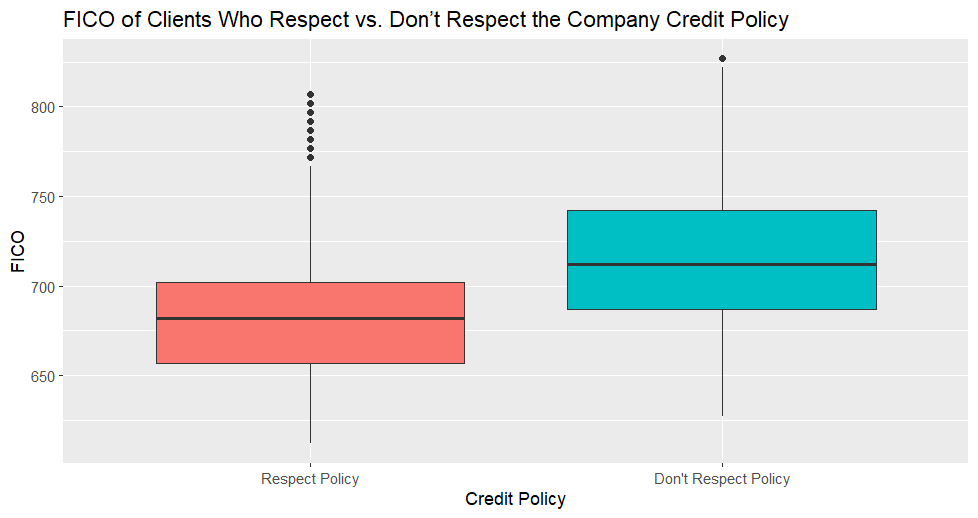
Descrizione generata automaticamente

Debt-Income Ratio is a measure that can give an idea of the stability of the potential borrower. The median is 12.72. The frequency sinks when the dti is bigger than 25.

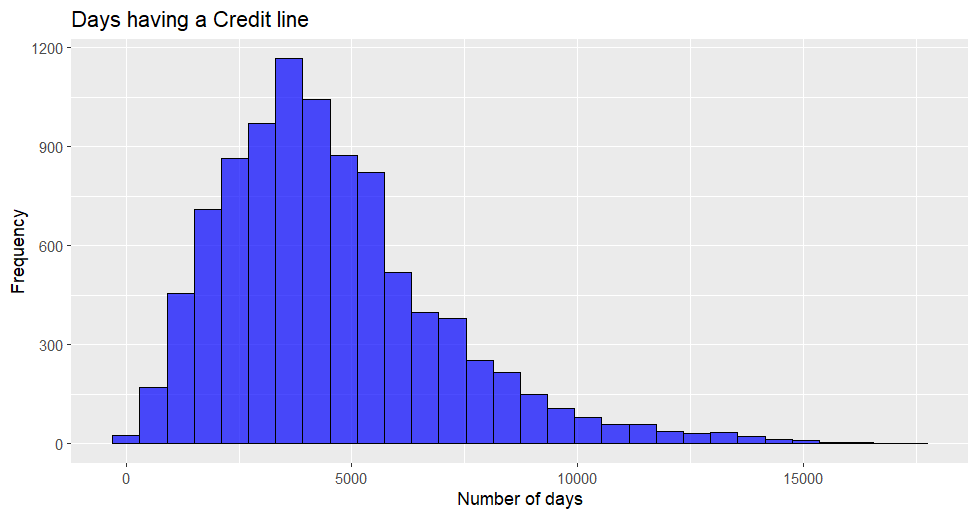
Immagine che contiene schermata, testo, diagramma, Diagramma

Descrizione generata automaticamente

The FICO Credit Score has a range of [612-827], with a median of 707.



Borrowers that respect the credit underwriting criteria have a lower credit score.



Days of Credit lines follow a right-skewed distribution, with a median of 4110.5.

Immagine che contiene testo, schermata, schermo, diagramma

Descrizione generata automaticamente

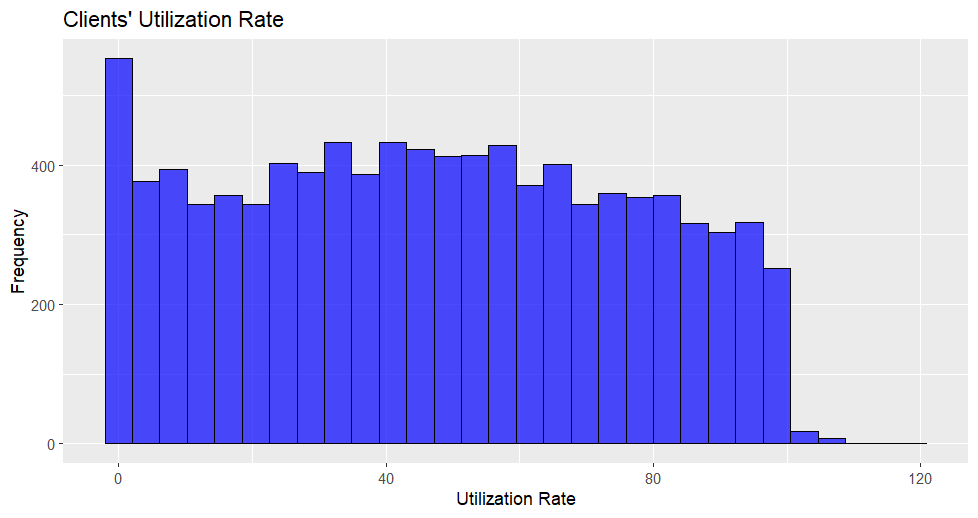
Most of the borrowers do not have a debt due to monlthy credit card payments.   
It’s iteresting to notice that the percentage of clients having 0 revolving balance within the two categories:

* Only 2.8% of clients that respect the credit criteria have a 0 revolving balance;
* The 5.2% of clients that respect the credit criteria have a 0 revolving balance.

Immagine che contiene testo, schermata, schermo, diagramma

Descrizione generata automaticamente

There’s not significant difference between the two category, concerning Revolving Balance.



The median utilisation rate is 46.2%.

Few borrowers have an utilisation rate lower than 100%

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

We have an higher utilisation rate for clients that respect the company’s credit criteria.

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

The 75% percent of the borrowers have a number of inquiries lower or equal to 2 in the last 6 months.

Immagine che contiene testo, schermata, numero, Diagramma

Descrizione generata automaticamente

Borrowers that respect the company’s credit policy have an higher number of inquiries.

Immagine che contiene testo, schermata, schermo, Diagramma

Descrizione generata automaticamente

Most of the borrowers have never been delinquent in the last 2 years.

Immagine che contiene testo, schermata, schermo, Diagramma

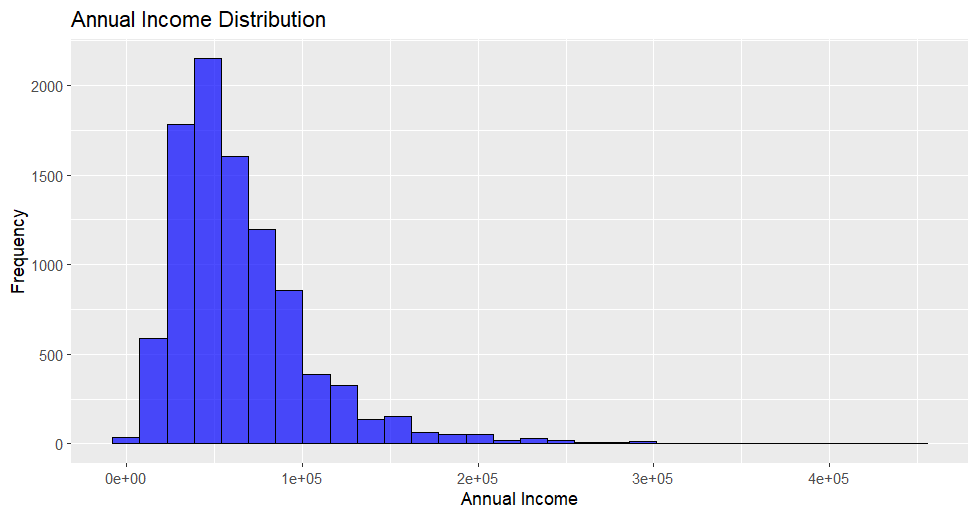
Descrizione generata automaticamente

Most of the borrowers have not derogatory public records (bankruptcy filings, tax liens, or judgments).

Immagine che contiene testo, schermata, Carattere, Diagramma

Descrizione generata automaticamente

Most of the clients have not fully paid the loan.



Annual income follow a right-skewed distribution with median equal to 55,764.

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

There is not a significant difference between the two groups reguarding Annual Income.

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

Debt follow a right-skewed distribution with median 666,795.7.

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

There’s not significant difference reguarding debt between the two groups.

Immagine che contiene testo, schermata, diagramma, Diagramma

Descrizione generata automaticamente

It appears that there's a relationship between income and maximum debt: as income increases, the maximum debt also tends to increase. There are instances where borrowers request amounts below this limit, indicating that they don't always seek the maximum loan they could qualify for.

4. Insights from the data

There are multiple variables used to measure the risk of a borrower defaulting on their loan. Some like the FICO credit score, utilization rate, number of inquiries, and the absence of monthly credit card unpaid debt, suggest that those not adhering to the company's credit policy present a lower risk. Furthermore, clients who abide by the policy are offered higher interest rates. This seems counterintuitive, as the credit policy is designed to categorize clients based on their financial stability and maximize the likelihood of repayment. These doubts highlight the need for further data analysis.

5. Ask Fase

The main goal of this analysis is to answer the following questions:

* What variables influence the company's decision on whether clients meet the company's credit policy or not?
* Which variables impact the decision on interest rates?
* Can borrowers be grouped (clustered) based on the available data before investors decide to lend them money? Are these clusters meaningful, and what are the significant differences between them?

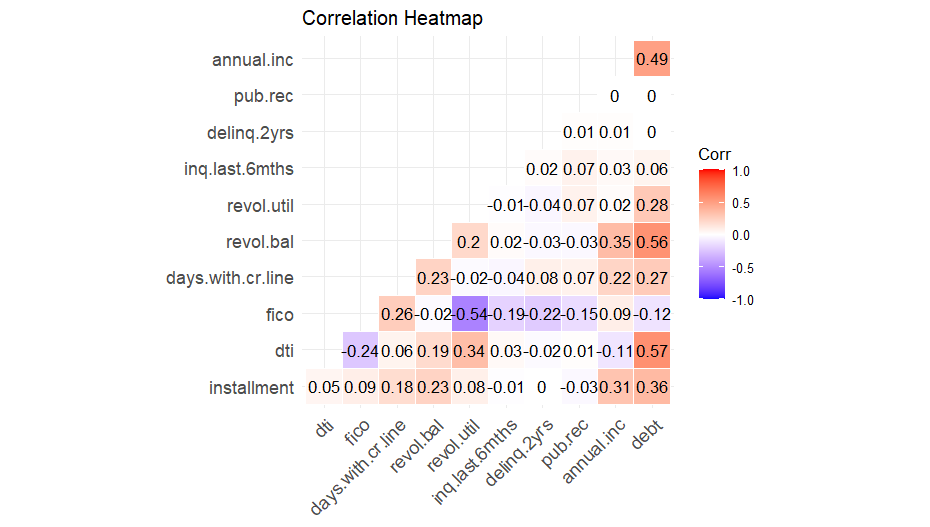
6. What are the factors that indicate a client is adhering to the company's credit policy?

To explore this, supervised learning techniques have been employed.

6.1 Column Selection

Given that the goal is trying to understand which variables make the company put its mark on the potential borrower, the not relevant variables have been excluded by the model: the interest rate, the total interest, and the fact that a borrower has repaid the loan are information not available when the borrower first approach the website. This because they are decided later, when the investor already selected the potential borrower.

The idea is to create a general linear model specifying the binomial family. This because we are modelling a binary variable that can’t be negative.

The following is a correlation heatmap, to check if there is a risk of multicollinearity between the independent variables of the model:

There is a high positive correlation between debt amount and debt to income ratio, debt amount and revolving balance, debt and annual income, and credit score and revolving line utilisation rate. This could cause multicollinearity.

To decide if the high correlation is a problem the Variance Inflation Factor (VIF) has been computed:

The VIF results for these variables are the following:

|  |  |
| --- | --- |
| Variable | VIF |
| FICO | 1.651582 |
| Annual Income | 3.343724 |
| Debt | 5.432976 |

The variable *debt* has a moderately high VIF, meaning that could cause multicollinearity.   
The decision is to keep it for the moment, trying to get as much information as possible from the dataset.

6.2 Simple Generalised Linear Model

The model is the following, and it includes all the variables selected in the *6.1* step:

Immagine che contiene testo, schermata, Carattere, documento

Descrizione generata automaticamente

The most significant variables are installment (positive coefficient), FICO Credit Score (positive coefficient), days with a credit line (with a positive coefficient), revolving balance (negative coefficient), the revolving line utilisation rate (positive coefficient), number of inquiries in the last 6 months (negative coefficient) and annual income (positive coefficient). When running the model, the RStudio compiler returned a warning: some variables may be deterministic, potentially leading to misleading results and higher coefficients. Comparing the distribution of the variables of the model in the two cases (respecting company policy or not), it has been possible to highlight the two deterministic variables: the number of inquiries in the last 6 months and the revolving balance.

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

After 9 inquiries, there are not records that respect company’s credit policy.

Immagine che contiene testo, schermata, diagramma, linea

Descrizione generata automaticamente

After a certain treshold of revoling balance, there are not records that respect company’s credit policy. One solution could be deleting the variables from the model, losing some information, or to use alternative models that shrink the coefficient of deterministic variables.

Before introducing these alternatives model, the current one has been tested, splitting the dataset in training set (80% of the records) and test set (20% of the records).

Then some indices for model evaluation has been computed.

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| Prediction\Reference | 0 | 1 |
| 0 | 233 | 47 |
| 1 | 136 | 1500 |

The model obtained 1550 true positives (TP), 233 true negatives (TN), 136 false positives (FP)and 47 false negative (FN).

The AUC (Area under the curve of the Receiver Operating Characteristic) is 0.8045, meaning that the model has a high discrimination power.   
The Accuracy is equal to 0.9044885.

The Precision is equal to 0.9168704.

The Recall is equal to 0.9696186.

The F1 score, a harmonic average between of Precision and Recall, is 0.9425071.

The model achieved excellent results, indicating that the variables used by the company to determine "company.policy" align closely with those included in the model.

6.3 Ridge Linear Regression

n