P02 - PD Models

ITESO, Universidad Jesuita de Guadalajara

Integrantes:
Andrés González Luna Díaz del Castillo
Susana Guadalupe Ascencio Martín
Luis Luengo Mayagoitia
Juan Pablo García Zaragoza
Santiago Riggen Romero

Profesor:
Luis Felipe Gomez Lopez

Executive Summary

This report presents the results of a credit modeling project that aimed to identify the best payers among a group of clients. The dataset's name is credit_risk_data_v2 and it contains a total of 74 data columns.

The analysis consisted of several stages, including exploratory data analysis (EDA), variable selection, predictor interpretation, model comparison, model selection, and explainability using stacking. The preprocessing steps started at first analyzing the type of data we have to know where we are standing with it, and, as we are doing the second project, most of it was about the data cleansing, such as eliminating the two columns with the identifiers because they won't be used and also the Nan values and missing data so it won't interfere with the analysis and also locating the null values.

After exploring the data, our objective was to identify five key predictors: loan amount, rate, annual income, age, and home ownership. We performed an in-depth analysis of these variables to gain insights into how they affect a client's ability to repay a loan.
Our analysis revealed that these variables have important implications for credit modeling and can help lenders make more informed decisions about which clients are likely to be good payers.

EDA

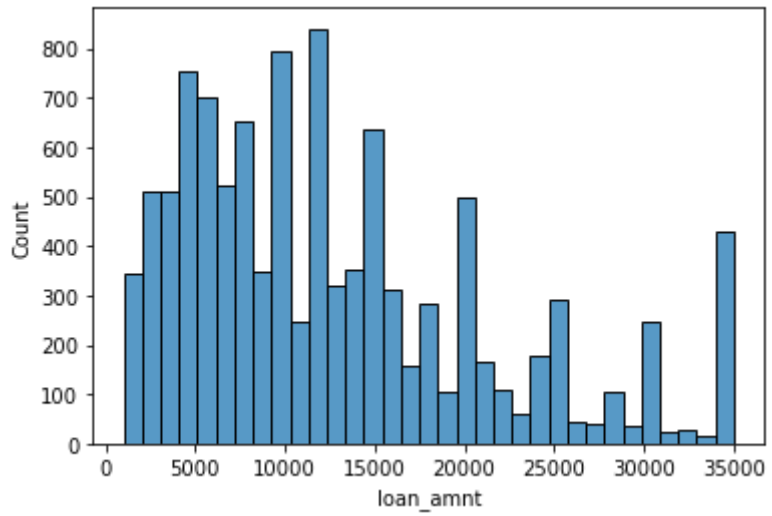
The Exploratory Data Analysis (EDA) is an important step in our analysis for summarizing data to gain insights and identify some patterns. EDA is particularly important in credit card analysis because it plays an important role in managing risk, detecting fraud, and optimizing customer experience. In this report, we will discuss the more relevant aspects for our project.

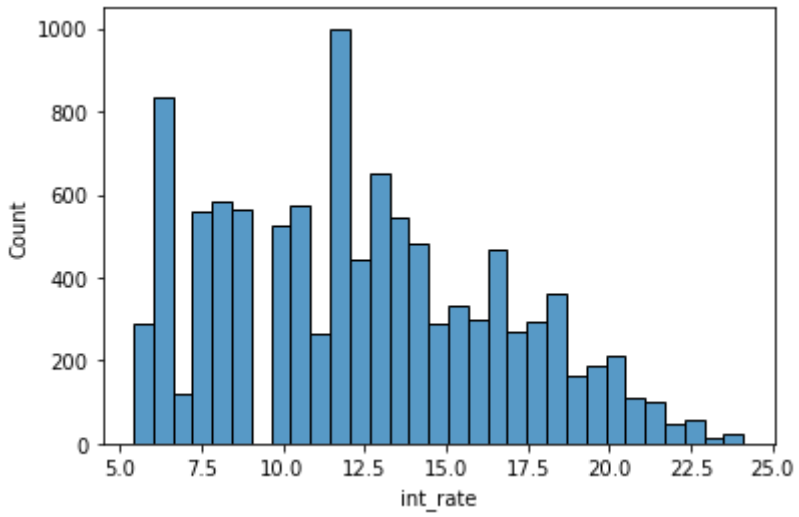
In this part of the project, we provide a detailed description of the dataset, including its size, shape, and data types. We perform an exploratory data analysis (EDA) to identify patterns, trends, and potential outliers. We also focus on the loan amount, rate, annual income, age, and home ownership columns and analyze their distributions, correlations, and other relevant characteristics.

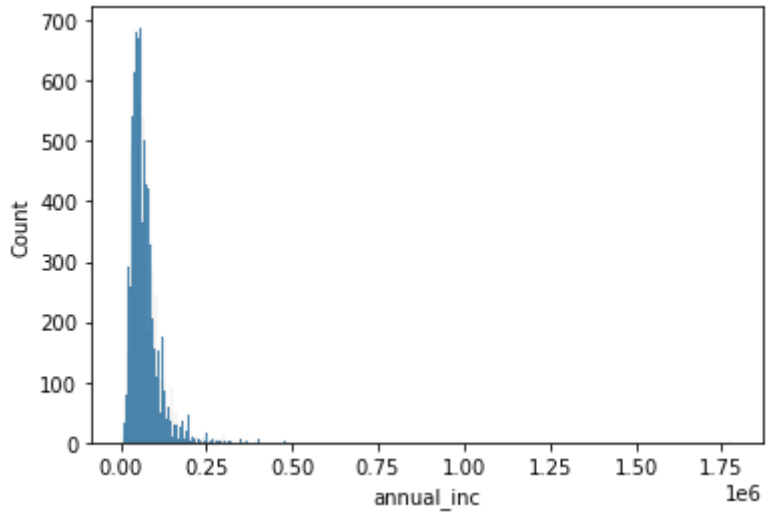
The first step in EDA is data collection, for this project data was collected from the source given and data is then stored and put into a structured format. Once collected, data needs to be cleaned and preprocessed. This involves removing duplicates, filling in missing values, and correcting any errors in the data. Data cleaning is crucial as it ensures that the analysis is based on accurate and reliable data.

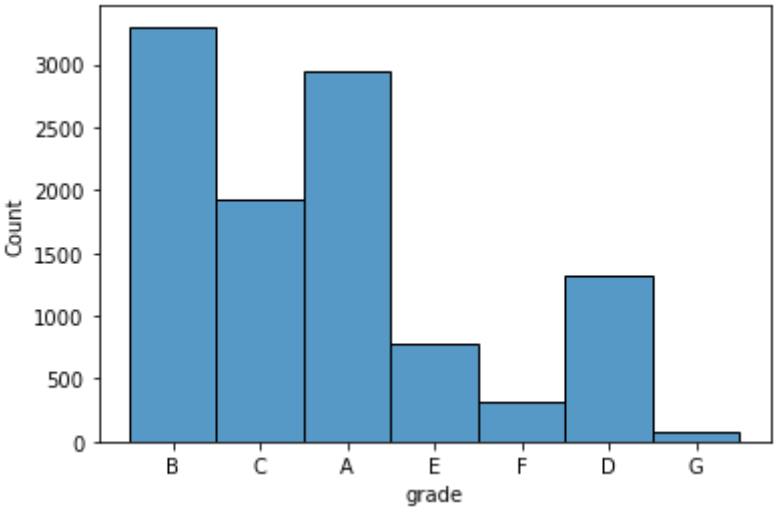
Then we get to data visualization. It's an important aspect of EDA. It involves creating visual representations of the data to identify patterns and trends for the selected variables, data visualization can be used to identify spending patterns, transaction volumes, and customer behavior. Visualization tools such as charts, graphs, and heat maps can be used to create visual representations of the data.

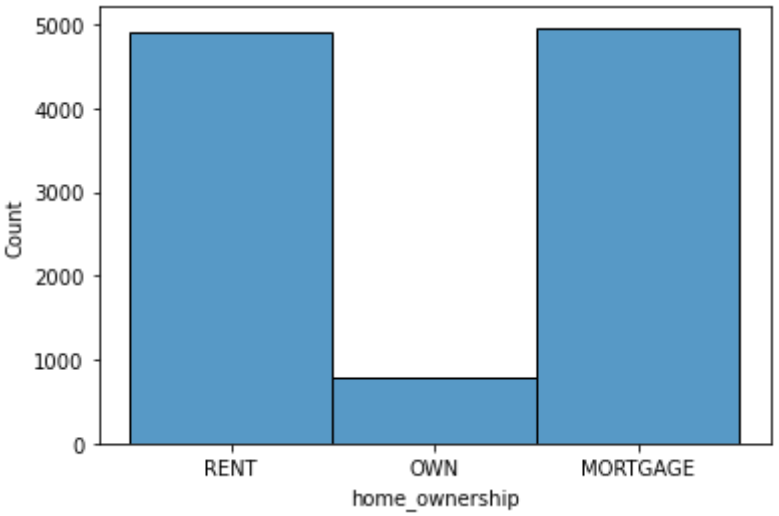
After that we start the statistical analysis which is another important aspect of EDA. It involves using statistical techniques such as regression analysis, hypothesis testing, and clustering analysis to identify patterns and relationships in the data. Statistical analysis can be used to predict customer behavior, identify risk factors, and detect fraud.






Count
700
600
500
400
300
200
100
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6
annual_inc


Count
3000
2500
2000
1500
1000
500
0
B
C
A
E
F
D
G
grade


Count
5000
4000
3000
2000
1000
0
RENT
OWN
MORTGAGE
home_ownership

Loan Amount

Loan amount is an important factor in determining a client's credit score because it reflects the client's ability to manage debt responsibly. A client who is able to borrow a significant amount of money and repay it on time and in full demonstrates a high level of financial responsibility, which is viewed positively by lenders. The credit financial institution sets a maximum limit of money, which the customer can use in part or in full, this is how we define this variable.

For the loan amount we analyze the distribution, the data information so we could detect outliers to take away unbiased data to see how important is the loan amount compared with several other variables that can help us predict how probable it will be for the client to pay or default. The bigger the loan amount is the more probable for the client to default

Interest Rate

An interest rate is the cost of borrowing money or the return on savings or investments. It is expressed as a percentage of the amount borrowed or invested and represents the cost of using someone else's money.

When it comes to credit, interest rates play a significant role in determining the cost of borrowing money. The higher the interest rate, the more expensive it is to borrow money, as borrowers have to pay back more than the original amount borrowed. The interest rate that a borrower is offered depends on various factors, including their credit score, income, debt-to-income ratio, and the type of loan they are applying for. The lower the interest rate is the cheaper it will be for the client to borrow the money.

Annual Income

The borrower's annual income provides a picture of their earning capacity and helps the lender determine the borrower's ability to make the monthly payments on the loan. Generally, the higher a borrower's income, the more likely they are to be able to repay their debt.

The variable talks about the total value of income or salary earned during a fiscal year. It refers to all earnings before any deductions are made, for it we analyzed its distribution to identify potential outliers, and be aware of how this variable affects a client's ability to repay a loan, the higher the annual income the client has the better score it will get. Annual income is one of the most critical factors in determining creditworthiness and will probably have a significant impact on a borrower's credit score.

Grade

This variable is already really important because it tells us how the client has behaved in the past with credits the clients have had. We use a client's intern credit grade as a tool to assess the risk associated with the possible loan. A higher credit grade indicates that the client is more likely to pay back the loan, while a lower credit grade indicates the opposite.

A good intern credit grade helps us locate a client way better so we can give them more favorable terms, while a poor credit grade can make it more difficult for a client to get approved for a loan or result in higher interest rates and less favorable terms.

Home Ownership

This variable talks about the status of the client of owning or not a property or a house, either outright or through a mortgage, where the homeowner has legal rights to occupy, use, and make alterations to the property.

Evaluating home ownership is important because it provides us with a clear understanding of the borrower's financial position and creditworthiness because we need to assess the borrower's ability to repay the loan, and owning a home can be a good indicator of financial stability, as it suggests that the borrower has a steady source of income and has been able to save up for a down payment

In some cases, home ownership can serve as collateral for the loan, which means that the lender has a form of security in case the borrower defaults on the loan.

Bins

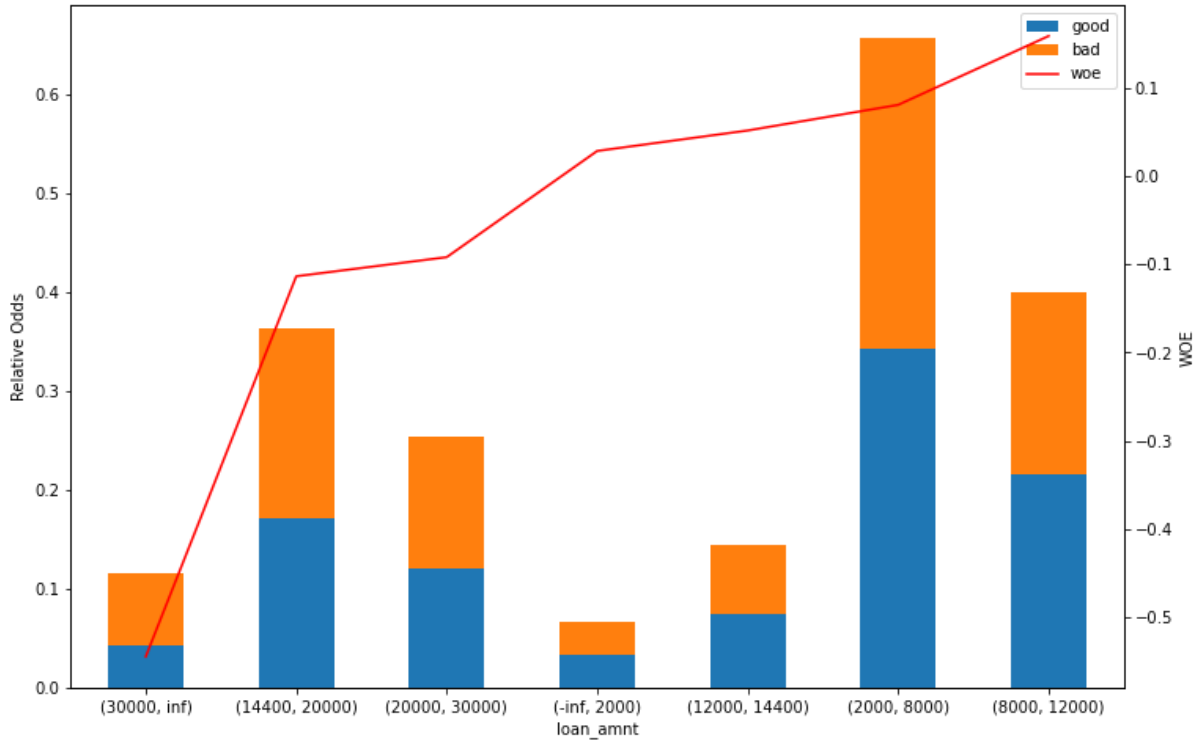
We used the process of “binning” which we can define as dividing a variable into a set of intervals or "bins", in this case it is helpful because the exact value of the variable is less important than its range or category. Binning, for this project, also helped us to reduce noise in the data and also deal with outliers.

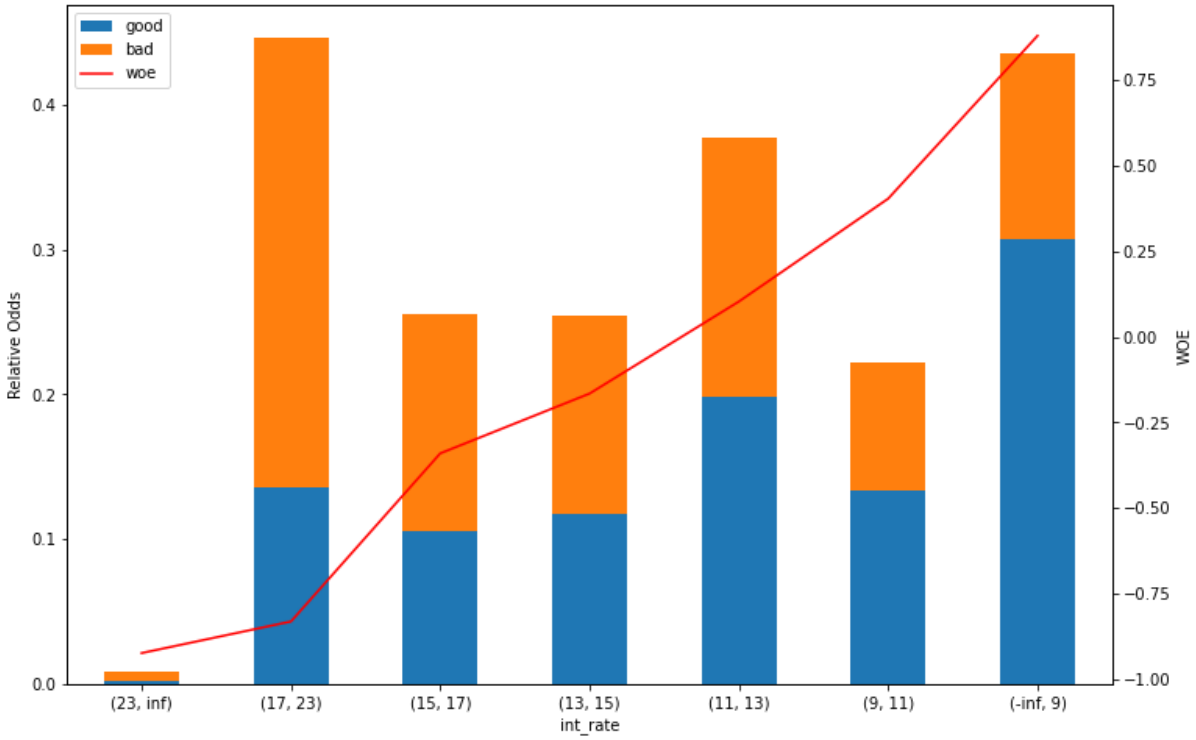
So we separated our variables to train in categories with this process which helped us to optimize and make more efficient having our variables this way to set aside the outliers in the dataset and, for our purposes, is way better to have segmented this type of variables that we’re using than having the exact data value of each file.

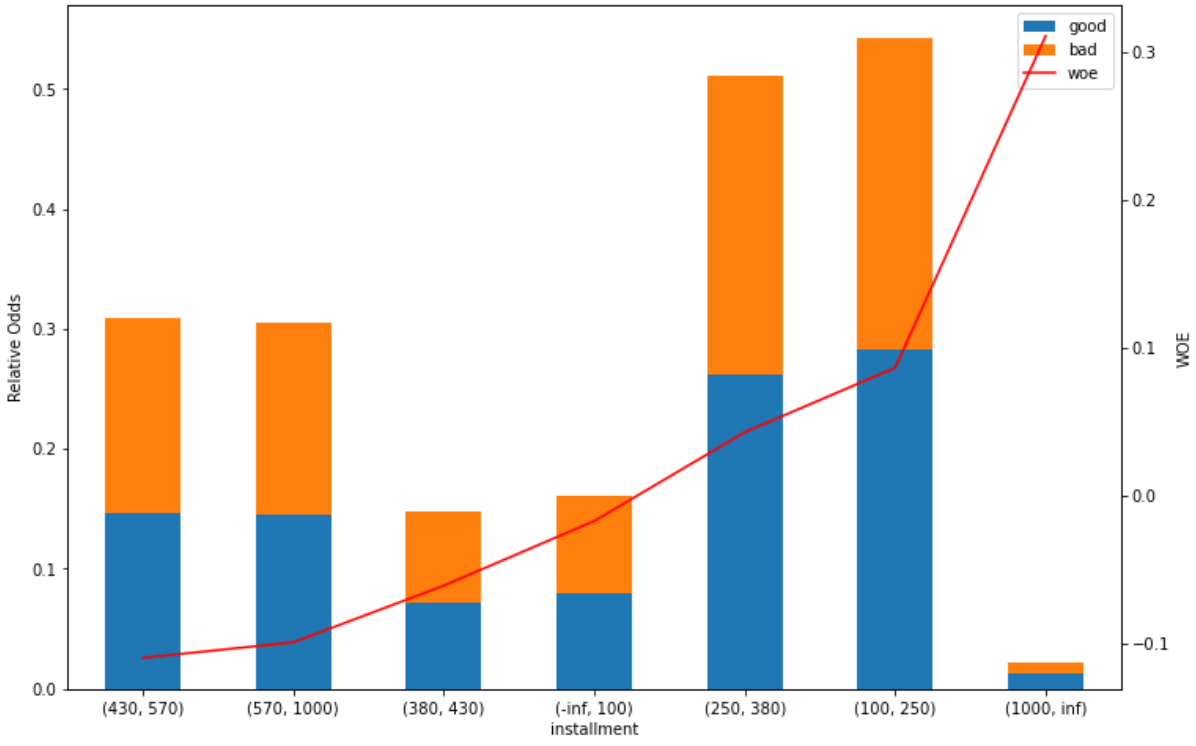
WOE

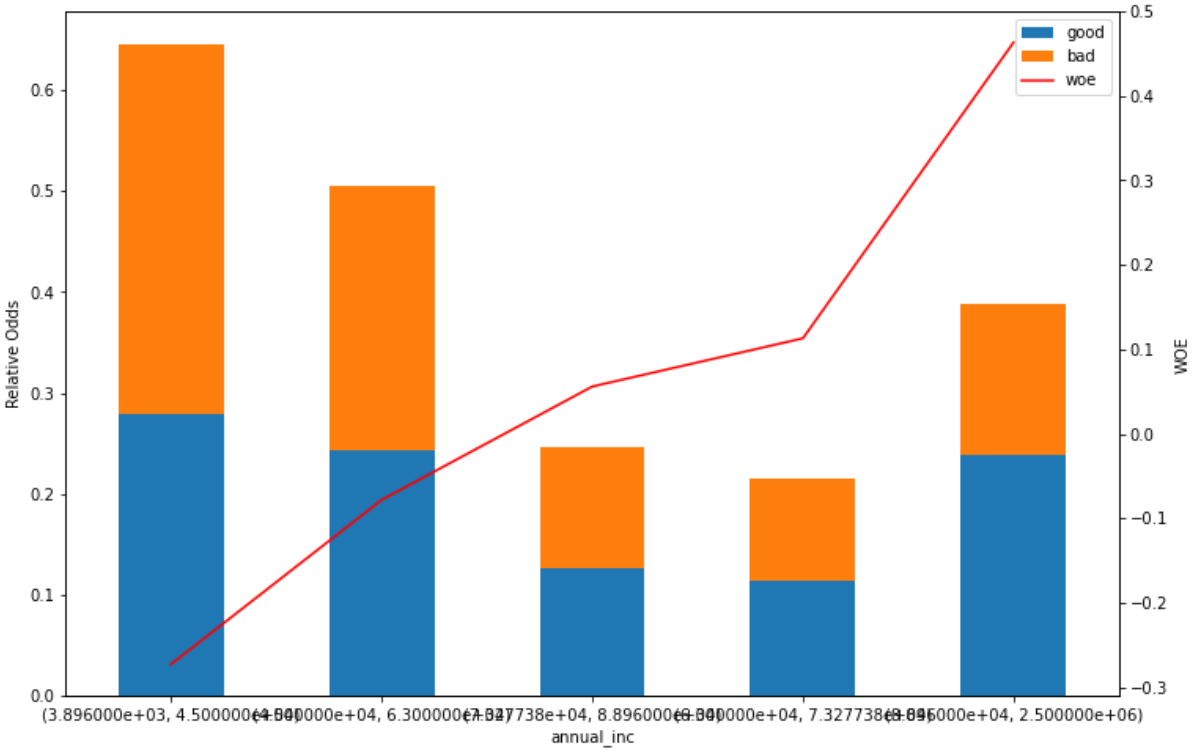
The WOE or Weight of Evidence helps to transform a continuous independent variable into a set of groups or bins based on similarity of dependent variable distribution. It tells the predictive power of an independent variable in relation to the dependent variable. Since it evolved from the credit scoring world, we can define it as a measure of the separation of good and bad customers. "Bad Customers" refers to the customers who defaulted on a loan. and "Good Customers" refers to the customers who paid back the loan.

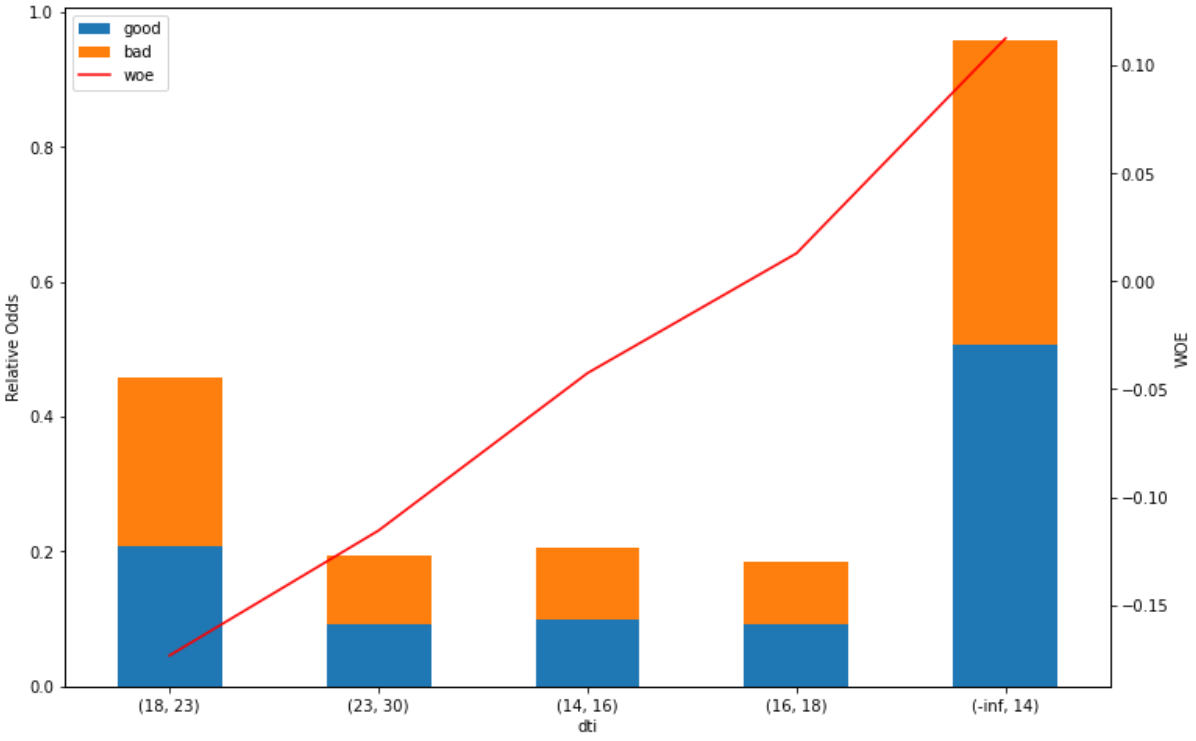
We can see The results of the transformation made with the WOE for the variables in the graphs:




good
bad
woe
Relative Odds
WOE
0.0
0.1
0.2
0.3
0.4
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
(23, inf)
(17, 23)
(15, 17)
(13, 15)
(11, 13)
(9, 11)
(-inf, 9)
int_rate

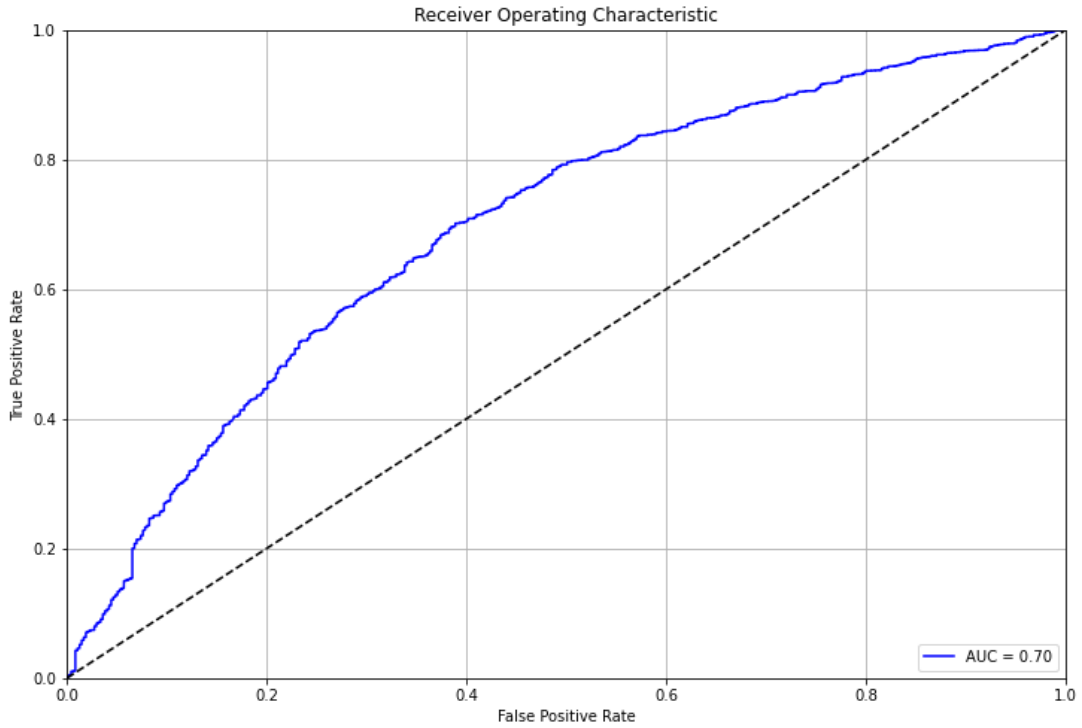

good
bad
woe
Relative Odds
WOE
0.0
0.1
0.2
0.3
0.4
0.5
0.3
0.2
0.1
0.0
−0.1
(430, 570)
(570, 1000)
(380, 430)
(-inf, 100)
(250, 380)
(100, 250)
(1000, inf)
installment


good
bad
woe
Relative Odds
WOE
annual_inc
(3.896000e+03, 4.500000e+04, 6.300000e+04, 8.896000e+04, 2.500000e+06)


good
bad
woe
Relative Odds
WOE
dti
(18, 23)
(23, 30)
(14, 16)
(16, 18)
(-inf, 14)

Logistic model

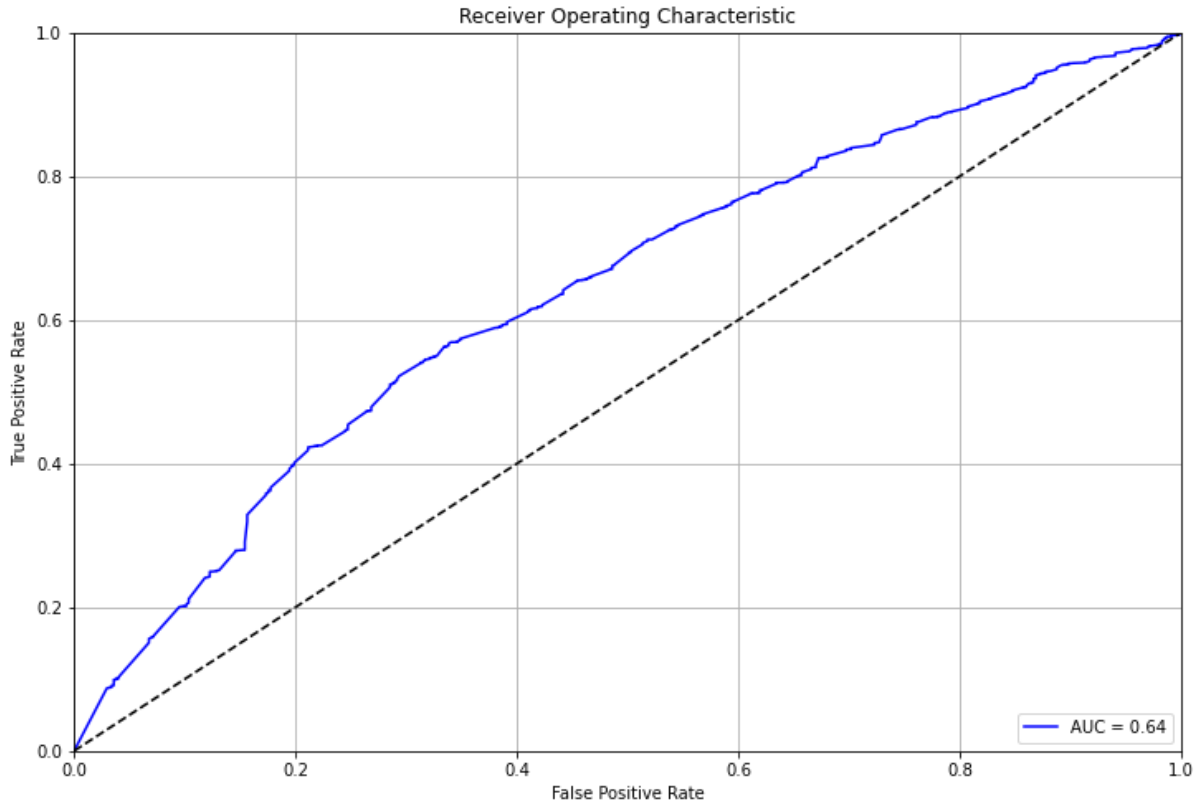
The logistic model is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more variables. This type of regression is an important tool in this case because it allows our algorithms used in machine learning applications to classify the data based on history. As additional relevant data comes in, the algorithms get better at predicting classifications within data sets.



ROC AUC: 0.69
F1: 0.92

Forest Classifier

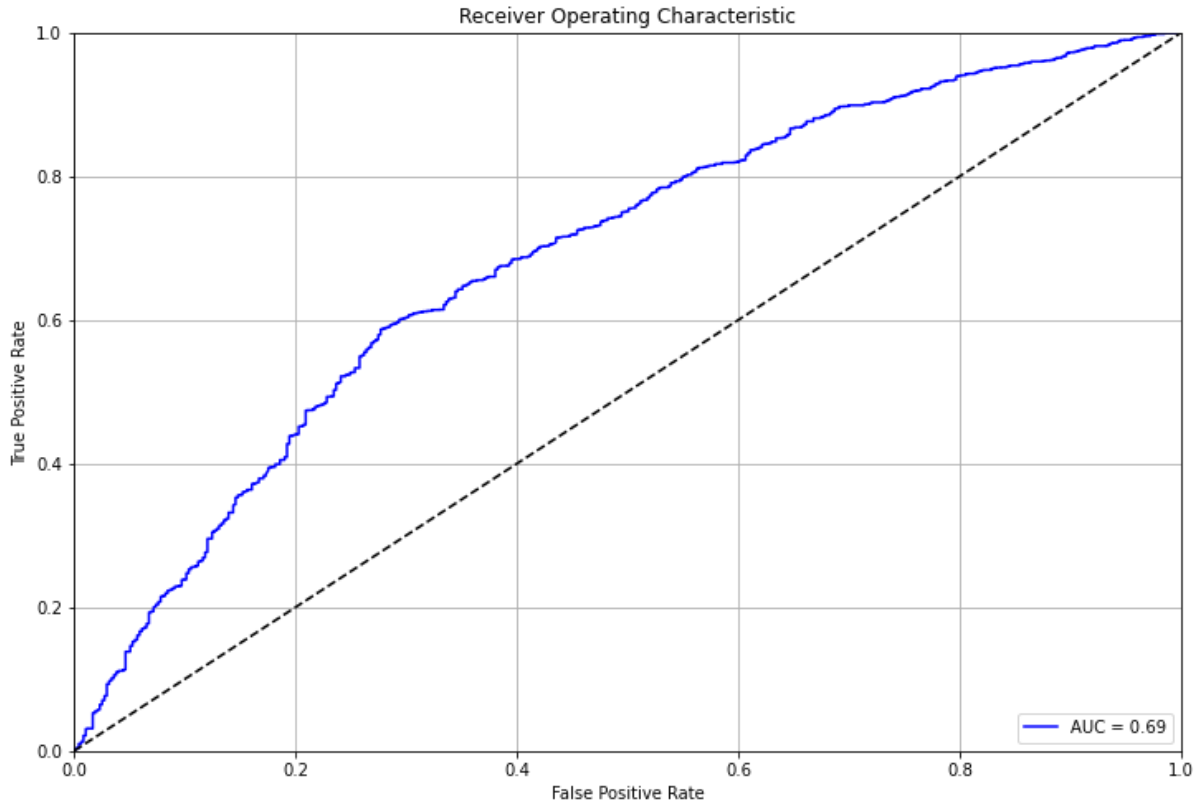
The ‘Forest Classifier’ or ‘Random Forest’ Random Forest is a classifier that contains a number of decision trees for the given dataset and takes the average to improve the predictive accuracy of that dataset
Multiple decision trees are created using different random subsets of the data. Each decision tree is like an expert, providing its opinion on how to classify it. Predictions are made by calculating the prediction for each decision tree, then taking the most popular result.



ROC AUC: 0.66
F1: 0.90

Gradient

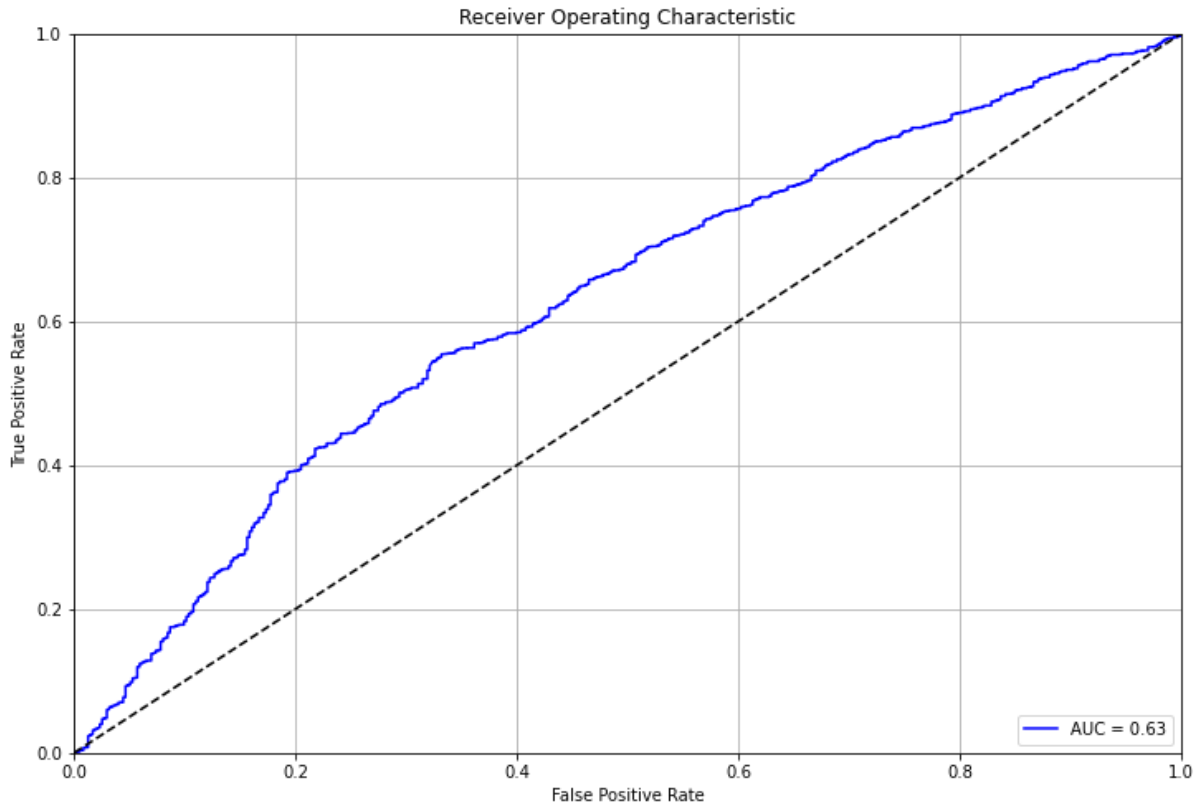
The gradient model is a technique which we used in regression and classification tasks. It lets us automate in a more efficient way, testing the machine learning model coefficients.
We can see in the graph that the gradient represents the slope where we're at, the more steep the more we want to get back so we can minimize the error



ROC AUC: 0.69
F1: 0.92

Stacking Regressor

The stacking regressor is a technique we also use in regression and classification tasks that helps us evaluate several models in a way that we get the result of a combination of the previous already shown models to get a new more optimal one.



ROC AUC: 0.63
F1: 0.90

Conclusion

By employing a gradient forest classifier, logistic model, and stacking regressor, we were able to explore different modeling techniques to better understand the relationships between the variables.
Given the results we got of each one of the models and also the results with the WOE, we can be certain that the models worked just fine but at the same time we should never forget that these are only predictors based on historical data and we can never be 100% certain that they will work with the same results or even similar in real life because there are lot of other variables that can change the output we get once we put it on practice in the real world but at the same time this type of analysis let us know how probable is something to happen and we can always keep on fixing the model to make it more and more accurate.