Irina Nobaeva

Loan Default Prediction Using Machine Learning

Lending loans is a significant source of revenue for banks, but it is not without risk. Loan defaults are a common occurrence and happen when a borrower fails to repay the money borrowed from a bank or lender. There are multiple reasons why people default on loans. Defaulting not only harms their credit score but also exposes them to potential legal actions and wage garnishment.

Organizations seek to predict loan default for consumer lending products. With access to historical client behavior data, they aim to identify higher-risk and lower-risk individuals among new consumers to assess the likelihood of default.

To address this issue, we can use machine learning techniques to develop a robust model that can predict whether a new borrower is likely to default on their loan. The banks have gathered extensive data on past borrowers, and we can help to create a powerful machine learning model for this purpose. In this project, I will build a classification model to predict weather the loan borrower will default or not.

The Loan Default dataset was retrieved from Kaggle website and comes in a csv format. The dataset provided is vast and includes various deterministic factors such as borrower's income, gender, loan purpose, and more.

Kaggle website

1. Data Wrangling

The raw data from Kaggle comprised of 148,670 Rows and 34 Columns. To enhance the efficiency of the model, I made careful decisions to exclude irrelevant columns that were

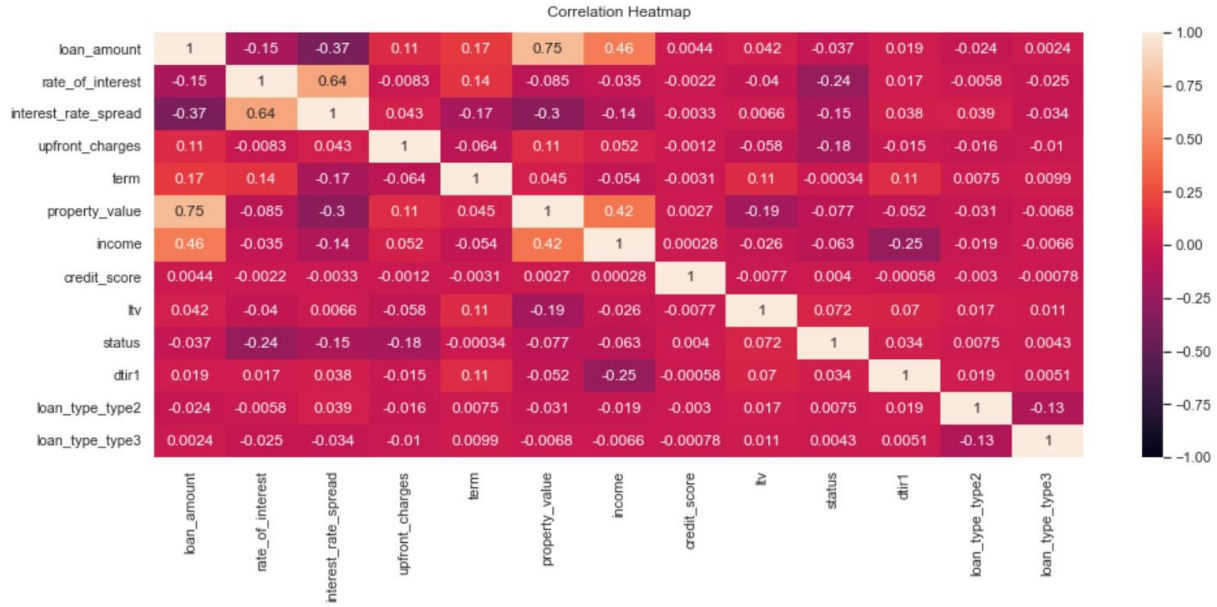
unlikely to significantly influence the outcomes. Additionally, I eliminated duplicate entries and filled in missing values.

During the analysis, I delved deeply into various application features including loan amount, age, interest rate, status, property value, and income.

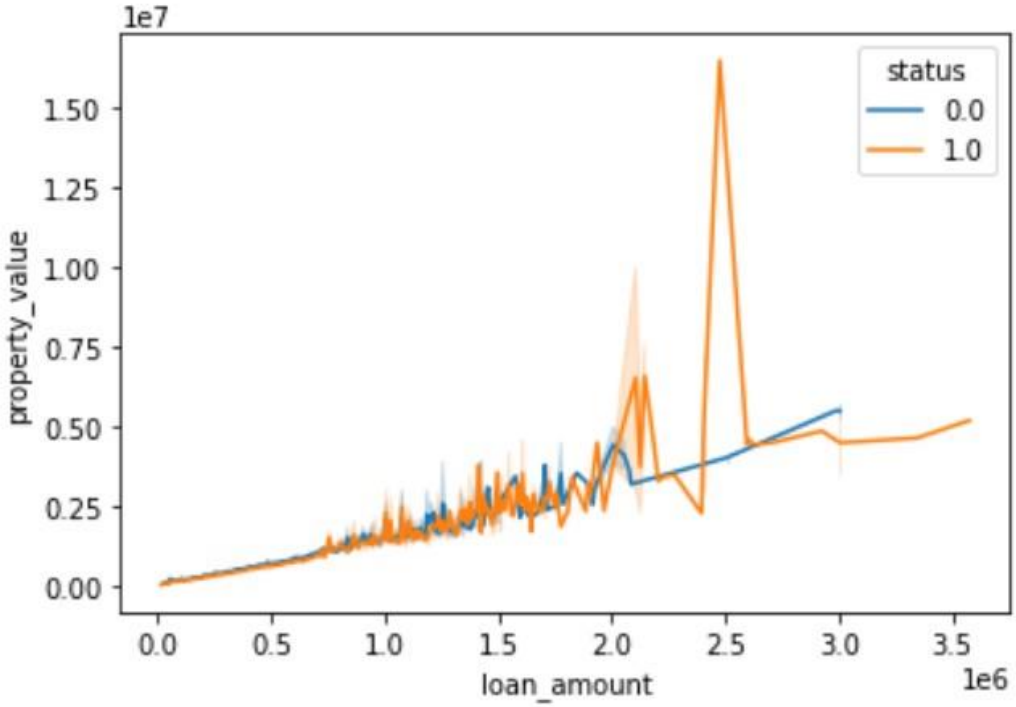
We also see that some of the features from the provided file are numeric and some of them are categorical.
Numerical data is the representation of measurable quantities of a phenomenon. We call numerical data "**quantitative data**" in data science because it describes the quantity of the object it represents.
Categorical data refers to the properties of a phenomenon that can be named. This involves describing the names or qualities of objects with words. Categorical data is referred to as "**qualitative data**" in data science since it describes the quality of the entity it represents.
So lastly, I divided the data into categorical and numerical values, handling them separately. All values were then converted into numerical formats to cater to our machine learning model.

2. Exploratory Data Analysis

With our data now in a pristine state, I proceeded with exploration and analysis. I started with plotting the pairwise relationships and creating a heatmap, to check out how our variables relate to one another.



The findings from the EDA analysis of the variables and correlations revealed that the strongest positive correlation ((r=0.75) is between the **Loan amount** and **Property value** variables. This makes sense, as typically the higher loan amounts tend to correspond to more expensive properties.



Several factors that can couse this correlation are:

- Market Demand and Property Prices: In areas where property prices are generally high due to strong demand or limited supply, borrowers may require larger loan amounts to afford properties in those markets.
- Property Evaluation: Property value is often assessed by considering factors such as location, size, amenities, condition, and market trends.
- Borrower Qualifications: Lenders evaluate borrowers' income, credit history, and debt-to-income ratio when determining loan amounts. Higher loan amounts may be approved for borrowers with higher incomes and better financial profiles, who are more likely to be able to afford properties with higher values.

Additionally, there was a moderately strong correlation between the **Rate of interest** and **Interest rate spread** (r=0.64). It appears that they are typically correlated because the interest rate spread is a measure of the difference between two interest

rates. The interest rate spread is calculated by subtracting a reference interest rate (such as a benchmark rate) from the actual interest rate charged by a lender.

There were two weak-moderately strength negative correlations. The first, between **Loan amount** and **Interest rate spread** (r=-0.37), what can be influenced by multiple factors like market conditions, the borrower's credit history, the loan's duration, and the lender's pricing strategy. Lenders may be more willing to provide competitive rates and narrower spreads for larger loans as they perceive them to be less risky due to factors like the borrower's creditworthiness, collateral, or the potential profitability of the loan.

The second, between **Income** and **Debt-to-Income Ratio (DTIR)** (r=-0.25), what can be caused because the DTIR is calculated by dividing an individual's monthly debt obligations by their monthly income. A lower DTIR indicates that a smaller portion of a person's income is allocated towards debt payments, which generally suggests a healthier financial position and better ability to manage debt. On the other hand, a higher DTIR indicates that a larger portion of income is allocated towards debt, which may suggest a higher risk of financial strain or difficulty in meeting debt obligations. Therefore, when income increases, assuming debt obligations remain constant, the DTIR decreases. This negative correlation means that as income rises, the proportion of income allocated to debt payments decreases, resulting in a lower DTIR. Conversely, if income decreases, the proportion of income allocated to debt payments increases, leading to a higher DTIR. It's important to note that this negative correlation assumes debt obligations remain constant. If a person takes on more debt without a corresponding increase in income, the DTIR could increase even if income remains the same or decreases.

3. Algorithms & Machine Learning

3.1. Preprocessing

The goal of the preprocessing work is to prepare our data for fitting models.

In the preprocessing step, we performed several tasks. First, I converted our categorical data into numerical format using the get_dummies() function in Section 3.2. This transformation allows us to work with categorical variables in our models effectively.

Afterwards, I proceeded to split our data into a training set and a test set. The train-test split is a common practice in prediction-based algorithms and applications to evaluate the performance of machine learning models. I allocated 70% of the data to the training set, which will be used to train our models. The remaining 30% was assigned to the test

set, which will be used to evaluate the performance of our trained models on unseen data.

With the completion of these preprocessing steps, our data is now ready for the modeling phase. The training set will be utilized to build and train various models, while the test set will enable us to assess their performance and generalization capabilities.

By appropriately preprocessing the data, we have laid the foundation for effective modeling and can proceed with training and evaluating our models.

3.2. Classification Models

For the final step I used four classification models:

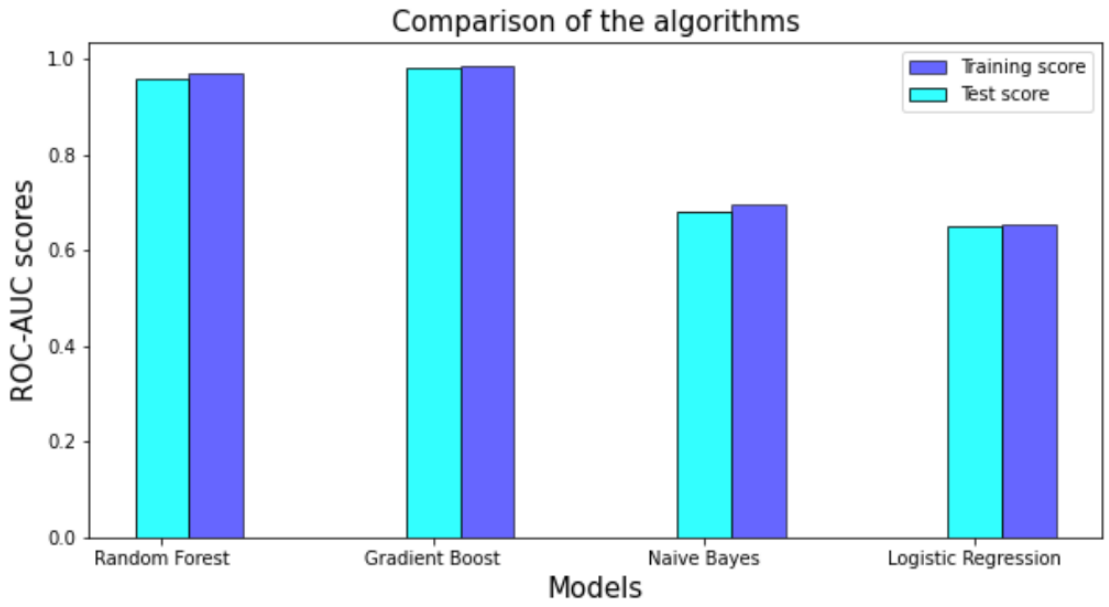
- Random Forest
- Naive Bayes
- Gradient Boost
- Logistic Regression

Models were employed and their performance assessed by analyzing the ROC-AUC scores on both the training and test data. The scores have been organized in a table and visualized through a plot.

```
              Algorithm              Model accuracy score
0         Random Forest              0.928676
1        Gradient Boost              0.963284
2           Naive Bayes              0.748974
3   Logistic Regression              0.755680
```

It is evident that the Gradient Boost and Random Forest models outperformed the others, showcasing remarkable results. Both models are ensembles, utilizing decision trees as their base.

	Algorithm	ROC-AUC train score	ROC-AUC test score
0	Random Forest	0.968779	0.957690
1	Gradient Boost	0.985129	0.982052
2	Naive Bayes	0.696242	0.680565
3	Logistic Regression	0.655198	0.649198



Subsequently, I conducted a thorough grid search and hyperparameter tuning specifically for Gradient Boost and Random Forest. This step consumed the most computation time, particularly for the Random Forest model. Using the optimized hyperparameters, I retrained the models and obtained separate predictions.

I evaluated the ROC-AUC scores using the optimized hyperparameters and observed a clear improvement in model performance. The final ROC-AUC scores for the Random Forest and Gradient Boost models were 0.97 and 0.98, respectively.

Future Research

The Loan Default Prediction models can be used in various ways to make informed decisions and manage risk effectively:

1. Loan Approval Process: Clients, such as banks or lending institutions, can integrate loan default prediction models into their approval process. By assessing the default risk associated with each loan application, they can make more accurate decisions about whether to approve or reject a loan. This helps mitigate the risk of lending to individuals or businesses with a high likelihood of default.
2. Interest Rate Determination: Loan default prediction models can aid in setting appropriate interest rates for borrowers. Clients can use the default risk assessment to adjust interest rates based on the perceived risk. Higher-risk

borrowers may be assigned higher interest rates to compensate for the increased default probability, while lower-risk borrowers may receive more favorable rates.

3. Portfolio Management: For clients with existing loan portfolios, loan default prediction models can assist in portfolio management. By continuously monitoring the default risk of individual loans within the portfolio, clients can identify potential defaulters early on and take proactive measures, such as offering refinancing options or implementing collection strategies, to minimize losses.
4. Fraud Detection: Loan default prediction models can also aid in identifying potential fraudulent loan applications. By analyzing various features and patterns within loan applications, such as inconsistencies in personal information or unusual financial behavior, clients can flag suspicious applications for further investigation, reducing the risk of approving loans based on fraudulent or misrepresented information.

It's important to note that loan default prediction models serve as decision support tools, and human judgment and expertise should still be considered when making lending decisions. These models provide insights and quantitative risk assessments that assist clients in managing their loan portfolios more effectively and reducing the potential impact of loan defaults.

For future research I would like to explore a few different areas like Feature Engineering, to investigate new and informative features that can capture borrower behavior and creditworthiness may improve default prediction models. This could involve exploring alternative data sources, such as social media data, transactional data, or alternative credit data, to extract valuable insights.

I would also try more Advanced Machine Learning Techniques, such as deep learning models or ensemble methods, could provide better predictive power and capture complex patterns in loan default data. Additionally, investigating the use of techniques like transfer learning or semi-supervised learning may help improve default prediction accuracy.

Online learning and real-time monitoring can also help in investigating the feasibility of implementing online learning techniques that can adapt and update models in real-time as new data becomes available. This would enable continuous monitoring of borrower behavior and enhance the responsiveness of loan default prediction systems.