Stijn

  stijnvanhijfte@hotmail.com

research report

Contents

[**Exploratory data analysis** 2](#_Toc197345462)

[**Model development** 2](#_Toc197345463)

[**Model performance & fairness evaluation** 2](#_Toc197345464)

[**Data limitations & Bias analysis** 3](#_Toc197345465)

[**EU AI Act** 3](#_Toc197345466)

[**Recommendations** 4](#_Toc197345467)

# **Exploratory data analysis**

The dataset structure consists out of approved loan applications – this might constitute bias (see later). There is a wide span of applicant profiles where we have data such as credit score, annual income, loan amount, and so on. When we have a look at the data itself, we see that there are some key outliers (i.e. high incomes or loan amounts). When we have a broader look at the distributions and averages, we tend to see relatively high credit score metrics (which makes sense as we only have approved loan applications). Therefore, we see a skew for credit scores towards the upper range. The income levels vary from modest to high, also here reflecting the approval of loans. When we compare the loan amount with the income, we see a moderate relation which might indicate limited debt-to-income ratio.

As we have a look at the correlations, we see a positive correlation between loan status and credit score (and also to a lesser extend with debt to income ratio, employment length and savings). We see negative correlations between income and payment to income ratio as well as loan term months (which can make perfect sense). There is also positive correlation between loan amount and loan to income ration, monthly payment and payment to income ratio. These relations all seem to make sense in a real world scenario (but in reality deeper analysis would help to determine whether we are missing crucial elements here). Additionally, this view has been created based on approved loans only, meaning we are missing part of the dataset.

# **Model development**

For the model development, we chose an interpretable model approach (vs an ensemble) as we want something that can explain how the model makes it decisions (see later in the document). Due to the nature of the project, we chose an alternative approach to arrive faster at a possible model. Through the PyCaret package, we are able to quickly evaluate different model families and see what model could be most relevant. We performed some preprocessing to ensure the data was ready to be used for ML models. When including all independent variables (except for the customer ID), we ended up with a logistic regression model and an accuracy of 62% (and AUC of about 50%).

However, based on experience, we know that we can improve the performance of the model as well as the explainability by reducing the number of features to a couple of key components relevant for the study. As such, we created a quick pipeline with a gradient boosting classifier to determine the relevance of the different independent variables for the creation of a model. When evaluating the variables, we determined the most relevant variables were credit score (by far the most influential), debt to income ratio, previous defaults and loan to income ratio.

Next, we performed the entire process with PyCaret again, but this time only including the key variables selected. Now we ended up with a linear discriminant analysis.

# **Model performance & fairness evaluation**

With the final model we developed, we achieved an accuracy of 76% and a AUC of 83%. When we did a cross-validation ourselves, we discovered a validation accuracy of 76% which was still significantly better than our original model. This means that this model is well performing even though this is misleading as the dataset used only contained approved loans.

If we wanted to go deeper into checking fairness of the model, we should be evaluating whether the model we created would discriminate against lower incomes or for example against people with a specific sex or background. The data provided might reflect underlying patterns which are not clear at the moment. As such, the model would have to be evaluated, tested and monitored for bias (from testing, and continuously during deployment). Similarly, the robustness of the model would have to be evaluated by stress testing it (for example low credit score range, high income, …). Based on the dataset provided, the model will probably be vulnerable to cases which it has never seen before.

# **Data limitations & Bias analysis**

A first element that immediately became clear was that there is clear selection bias in the dataset: there are only the approved loans available, where we have no view on the rejected applications. The model which was created for this project was as such based on biased data and cannot be generalized to the broader population of applicants for the bank. The predictions of this model for applications which have characteristics which fall outside of the dataset are unreliable.

A second element is historical bias which might exist as this dataset reflects the decision pattern of the bank in the past. Any biases which might exist in the human decision process for an application are now imbedded within the model (algorithmic bias might be a concern).

Next, we have no clear idea of the origin of the dataset and its quality. With 50,000 records, sample bias might be a concern (geography, customer demographics, …). At this point of time there is no way we can evaluate if the dataset can be trusted for its purposes.

Finally, the target variable is not what the bank ultimately wants to measure. The approval of a loan doesn’t mean that the risk of the loan is well understood. A loan can be approved but be unacceptable under certain circumstances and not default due to specific circumstances. “Default” is also a term with a wide range of definitions (90 days delinquent, charge-off, …).

Based on these considerations, the model cannot and should not be deployed to be used to help determine loan approvals. Underperformance and discrimination are a real possibility and due to this, more research would be needed.

# **EU AI Act**

Based on the definition in the EU AI Act, an automated loan approval system would be classified as a high-risk AI system. This creates requirements on risk, data quality, creditworthiness, oversight, transparency and non-discrimination.

The model would require a continuous risk management process where foreseeable risks are documented and mitigations are considered and put in place. Testing and validation of a model against “extreme” cases would already help to establish the possible risks which exist in the system. Additionally, the EU AI Act talks about systems which need to be accurate, resilient and secure throughout the entire lifecycle. In our case, this would mean continuous testing of the model and making sure it remains economically relevant. In case of regression, the model would either have to be updated or taken out of production until a new solution is available.

Another element which is emphasized by the EU AI Act is data quality and governance. An issue which exists within our current dataset is that it is not fully representative of all applicants and therefore is not statistically adequate for the task at hand (article 10). Governance would also mean proper documentation and traceability of data and model decisions, where we currently have no idea of how these might be implemented. To help deal with bias, synthetic data might be a correction we could apply.

Transparency and explainability are other elements which are highlighted in the regulation where applicants should understand the basis of automated decisions (something we also find back in GDPR). Interpretable models are one way of achieving this goal.

High-risk AI models also require human oversight and the ability for human intervention. If the proper model would be put into practice, it would only serve as a decision support tool for loan officers, so that a human-in-the-loop approach is the best approach. Borderline cases as well as rejected applications should be reviewed by a human being. On a higher level, an AI Review board could periodically review decisions by the system and check whether these are in line with expectations.

Non-discrimination and fairness are other elements which are mentioned for high-risk AI systems (minimize bias and avoid discriminatory outcomes). Items to achieve this could be data review before the model training, model bias testing and post-model audits.

# **Recommendations**

Based on what we have seen in our short research, here we offer a clear overview of our recommendations:

1. Improve the dataset: make it representative of all applications and perform data quality checks to ensure that the dataset is consistent.
2. Model performance monitoring: the model should be continuously monitored to ensure it remains relevant and takes clear decisions which are in line with the policies of the bank.
3. Data and AI governance: the bank should have a clear framework in place to ensure data lineage, use and quality. On top of this, the model should be reviewed by an oversight committee and to ensure continuous compliance with regulations. Training of people within the bank is an essential element to achieve this goal.
4. Fairness and compliance auditing: the model should be continuously tested to ensure the decisions remain fair. If certain groups are rejected more than others (racial discrimination, regions, …), there should be clear decisions taken to update the system accordingly. A security and privacy assessment could help hear as well to ensure no customer data is leaked which is identifiable.
5. Human-in-the-loop use: this cannot be deployed as a fully automated system and a proper human support factor should be implemented in the overall framework.
6. Ongoing model improvements: when the model is deployed, there should be continuous improvements so that it remains relevant in a changing market, with changing bank policies and adapting customers
7. Customer communication: finally, the bank should clearly communicate with customers that it is making use of the model. This is in line with the EU AI Act which focuses on transparency as well and as such becomes a legal requirement.