Data Janitors

Work Integrated Learning Project

GROUP 8

Alex Peter Thomas	(S3925735)
Isxaq Warsame	(S3658179)
Simran Sidhanti	(S3940756)
John Fergus Murrowood	(S3923075)
Matthew John Bentham	(S3923076)
Udit Dinesh	(S3879492)

October 21, 2022

TABLE OF CONTENTS

1. Introduction

Nowadays, with the ever-increasing number of machine learning algorithms developed and available, the accuracy gained has drastically improved for various models, but along with this improvement the models have slowly morphed into a Blackbox for the users. Predictions and insights gained from such 'Blackbox models' become insignificant if we cannot provide the reasonings and explain how the model ascertained its outcomes [11]. Thus, it is imperative to have better explainability of the model to understand why certain decisions were taken for a given prediction.

In the financial sector, various banks use prediction models to determine the credit risk of various customers based on previously gathered client data coupled with risk models along with various other inputs. These inputs are then utilised in a model with the aim of determining the risk profile of a given client and subsequently whether to grant them a loan or accept their credit application. The average duration for a home loan approval in Australia takes about 4-6 weeks [9], with the application phase taking about a week after submission. Therefore, customers are often left in the dark about the process and are only given an approval or denial with no explanation of how this decision was arrived at. Having information on why the decision was made is vital in aiding the individual to become better prepared in their next application. A further benefit of improving the transparency of the models behind this decision is that it improves the relationship between the client and the lender as clients are more willing to be understanding of the decisions taken if they are explained well.

Our Data Janitors website aims to provide customers and small-scale lenders the ability to upload credit profiles and evaluate their credit risk by using an efficient machine learning prediction model. It will also highlight and explain the major contributing factors for the given decision, why the given decision was predicted and for the users with elevated risk ratings, various mitigation plans can be offered to reduce the risk they pose to a business.

2. Problem Definition

As very complex machine learning models start to become widely adopted in the finance sector, especially for customer and risk management uses, due to their high predicative capabilities, the ethical dilemmas that arise because of automating important decisions such as rejecting someone's loan is often overlooked. Because of this, consumers and regulators are given no insight into the decision-making process of these automated decisions which opens the door for inequitable and biased models and overall unethical immoral business practices.

Additionally, as large banks start collecting more and more data, small creditors and loan providers lack the sufficient resources and data availability to compete with these corporations in terms of the use of machine learning models to provided adequate customer and risk management for their business. Similarly, they are unable to compete with larger lenders and creditors in having a sizeable workforce to have a person look through and evaluate many applications, therefore automation is required for smaller lenders to take on more clients without taking on too much risk or without properly assessing potential clients. Without a stuffiness open-source model to help bridge the gap, small businesses are forced to either adopted a much larger amount of risk with their client base or refuse business to a much larger proportion of their clients. Having an open-source model is also valuable for the use of loan applicants to be able to put in their details to see the likelihood of a loan before they apply in order to see if they may be successful or what they may have to change in order to be successful for a future application.

3. Methodology

3.1 Data Collection

The Data is sourced from a machine learning challenge setup by the fico community with an aim to find better explainability which will help data scientists understand their datasets and the models' predictions of financial risk assessment better, also uncover and check for biases, and ultimately create clear models.

Challenge Link: https://community.fico.com/s/explainable-machine-learning-challenge (Links to an external site.)
Dataset: https://drive.google.com/drive/folders/1SCreh1F12HDJx1vuw9Xpm3QVMS_UBmzH

3.2 Data Processing and Feature Engineering

Data Retrieval

The data was extracted from the linked dataset [4]. The data consists of 23 features each with its own significance.
The explanations for each of the features are as below:

Feature	Description
MSinceOldestTradeOpen	Months since oldest approved credit agreement
MSinceMostRecentTradeOpen	Months since last approved credit agreement
AverageMInFile	Average Months in File
NumSatisfactoryTrades	Number of credit agreements on the customer's credit bureau report with on-time payments
PercentTradesNeverDelq	Percentage of credit agreements on the customer's credit bureau report with on-time payments
NumTotalTrades	Total number of credit agreements the customer has made
PercentInstallTrades	Percentage of instalment trades the customer has

MSinceMostRecentInqexcl7days	Months since most recent credit inquiry into the customer's credit history (excluding the last 7 days)
NetFractionRevolvingBurden	Customer’s revolving burden (portion of credit card spending that goes unpaid at the end of a billing cycle/credit limit)
NetFractionInstallBurden	Customers instalment burden (portion of loan that goes unpaid at the end of a billing cycle/monthly instalment to be paid)
PercentTradesWBalance	Number of trades currently not fully paid off by the customer
ExternalRiskEstimate	Consolidated/Combined version of the risk markers
NumTrades60Ever2DerogPubRec	Number of trades above 60
NumTrades90Ever2DerogPubRec	Number of trades above 90
MSinceMostRecentDelq	Months since the most recent payment past the due
MaxDelq2PublicRecLast12M	Max number of payments past their due date. (Last 12 months)
MaxDelqEver	Max number of payments past their due date.
NumTotalTrades	Total number of trades
NumTradesOpeninLast12M	Total number of trades open in the last 12 months
MSinceMostRecentInqexcl7days	Months since a credit inquiry was conducted (excluding 7 days)
NumInqLast6M	Total number of credit inquiries conducted in 6 months
NumInqLast6Mexcl7days	Number of revolving trades with balance
NumRevolvingTradesWBalance	Number of instalment trades with balance
NumBank2NatlTradesWHighUtilization	Number of banks to National trades with a high utilization ratio
RiskPerformance	Calculated credit risk of the profile (Target variable)

Data pre-processing

To convert the raw financial data into a more useful and efficient format for our machine learning various pre-processing tasks need to be completed. These steps included:

1. One-hot encoding of all special values in each column (-7,-8,-9)
2. Min-max scaling all numerical features to reduce the effect of noise in the dataset
3. Removal of outliers and missing values
4. Splitting the Dataset into Train and Test sets

Feature Selection & Hyperparameter tuning

The importance of the aforementioned features was identified through the use of the Random Forest Classifier. This enabled us to determine the features most relevant to our classification problem as seen below:

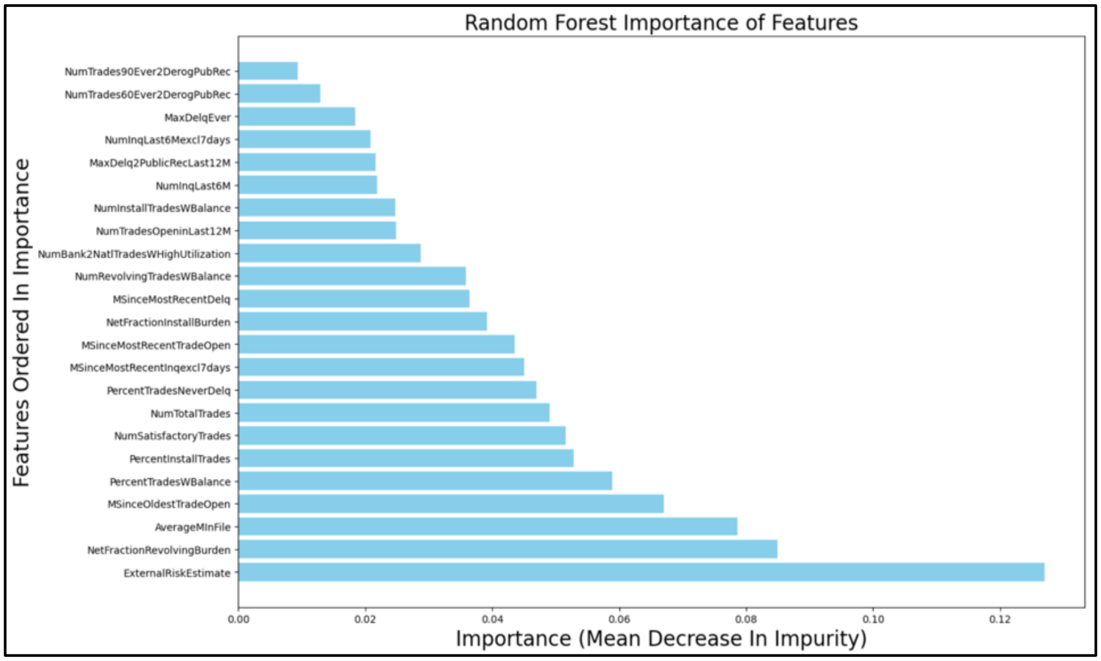


Fig 1. Random Forest importance of features

As per the Random Forest Classifier Feature importance the feature ExternalRiskEstimate was found to be the most relevant variable in comparison with the target variable. For all tested models below, to reduce the number of input variables and therefore computational complexity of each model and to remove redundant or irrelevant features for our problem, we used these random forest feature importance values in a grid search to determine the number of important features needed to get the best result on the test set. Additionally, to further fine-tune the overall model, hyperpresence tuning was also performed in a pipeline with feature selection using stratified 5-fold cross validation.

3.3 Data Classification Models

For the classification of the model, we have mainly worked with three explainable classification models a Decision Tree Classifier, LightGBM and a Logistic Regression binary classification model.

Decision Tree Classifier:

A Decision Tree is a supervised Machine learning algorithm wherein multiple rules are used to make decisions [3]. Utilizing the dataset features to create a series of yes/no questions and continually splitting the dataset until we isolate all datapoints belonging to each class.

Using the Decision Tree classifier, we will be able to depict to the end user the path or the reason why the decision is made. This will lead to clear understanding for the user on why the application was unsuccessful. If possible, they can work on improving those aspects in their next application process.

The decision tree model is very easily made explainable to any user of the website as each step and decision the tree made in arriving at its prediction can be seen, an example of the tree used for the dataset can be seen in figure **x** where all steps taken can be seen. Therefore, it is easy to see what aspects of a person's details may result in a risky prediction and give potential directions for people applying for credit areas where they can improve.

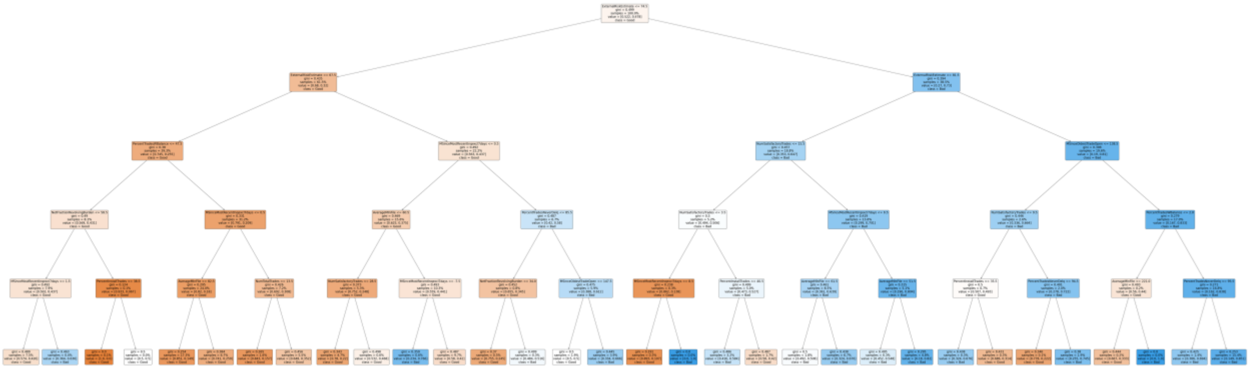


Fig 2. Decision Tree

LightGBM Model:

LightGBM, an open-source algorithm, utilises a tree-based learning algorithm and is a gradient boosting framework [10]. It is more efficient and distributed compared to other boosting algorithms. LightGBM can handle large amount of data, with low memory use, it has good accuracy and faster training speeds. The main difference between this and other models is that LightGBM grows leaf wise and other algorithms grow level wise, this means it builds trees in the vertical direction rather than in the usual horizontal direction.

Gradient boosted machines have many advantages over more simple decision tree algorithms, for instance it generally results in a higher accuracy, this is because they are able to build many smaller weaker learning tree models and continually improve them through boosting, until they achieve a greater accuracy than a simple decision tree [6]. However, a disadvantage of this is the LightGBM tree is less explainable than a simple decision tree model therefore, LIME was used to make the models predictions more interpretable. An example of the lime output can be seen below in figure 4. LIME works by modifying a single model input in different ways and each time seeing what effect the modification of that point has on the prediction and is thus able to work out what features the model uses to make its predictions. Lime then outputs a list giving the contribution of each feature of a point in the prediction of the model, thus telling potential clients why they have been predicted as risky by the model. This provides invaluable insight into the model however we cannot see the exact paths the model has taken when compared to the simpler decision tree, instead we get more of an approximation.

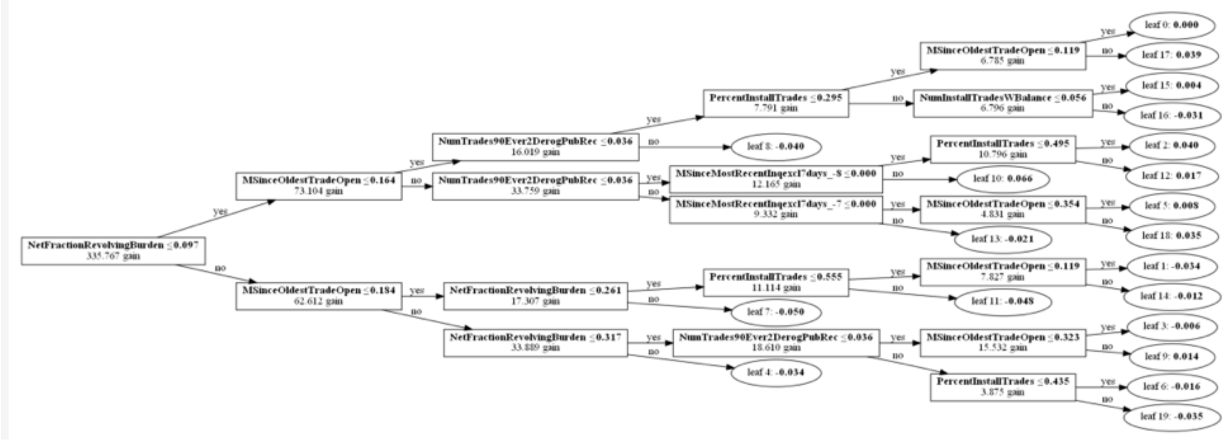
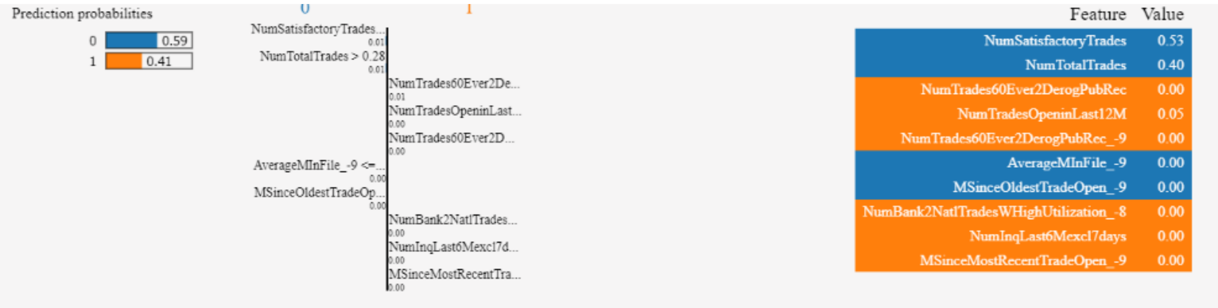

NetFractionRevolvingBurden ≤0.097
335.767 gain
yes
no
MSinceOldestTradeOpen ≤0.164
73.104 gain
MSinceOldestTradeOpen ≤0.184
62.612 gain
NumTrades90Ever2DerogPubRec ≤0.036
16.019 gain
NumTrades90Ever2DerogPubRec ≤0.036
33.759 gain
NetFractionRevolvingBurden ≤0.261
17.307 gain
NetFractionRevolvingBurden ≤0.317
33.889 gain
PercentInstallTrades ≤0.295
7.791 gain
leaf 8: -0.040
MSinceMostRecentInqexcl7days_-8 ≤0.000
12.165 gain
MSinceMostRecentInqexcl7days_-7 ≤0.000
9.332 gain
PercentInstallTrades ≤0.555
11.114 gain
leaf 7: -0.050
NumTrades90Ever2DerogPubRec ≤0.036
18.610 gain
leaf 4: -0.034
MSinceOldestTradeOpen ≤0.119
6.785 gain
NumInstallTradesWBalance ≤0.056
6.796 gain
PercentInstallTrades ≤0.495
10.796 gain
leaf 10: 0.066
MSinceOldestTradeOpen ≤0.354
4.831 gain
leaf 13: -0.021
MSinceOldestTradeOpen ≤0.119
7.827 gain
leaf 11: -0.048
MSinceOldestTradeOpen ≤0.323
15.532 gain
PercentInstallTrades ≤0.435
3.875 gain
leaf 0: 0.000
leaf 17: 0.039
leaf 15: 0.004
leaf 16: -0.031
leaf 2: 0.040
leaf 12: 0.017
leaf 5: 0.008
leaf 18: 0.035
leaf 1: -0.034
leaf 14: -0.012
leaf 3: -0.006
leaf 9: 0.014
leaf 6: -0.016
leaf 19: -0.035

Fig 3. LightGBM Decision Tree



Feature	Value
NumSatisfactoryTrades	0.53
NumTotalTrades	0.40
NumTrades60Ever2DerogPubRec	0.00
NumTradesOpeninLast12M	0.05
NumTrades60Ever2DerogPubRec_-9	0.00
AverageMInFile_-9	0.00
MSinceOldestTradeOpen_-9	0.00
NumBank2NatlTradesWHighUtilization_-8	0.00
NumInqLast6Mexcl7days	0.00
MSinceMostRecentTradeOpen_-9	0.00

Fig 4. LIME

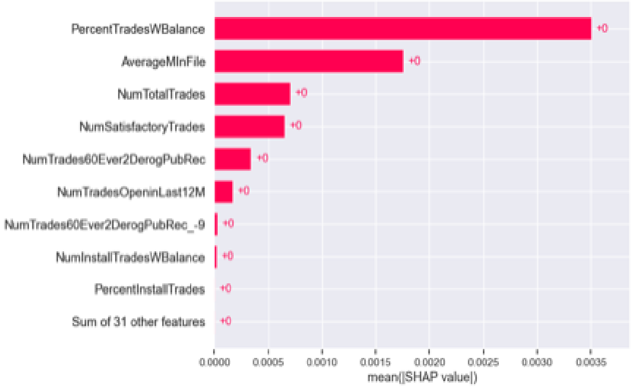
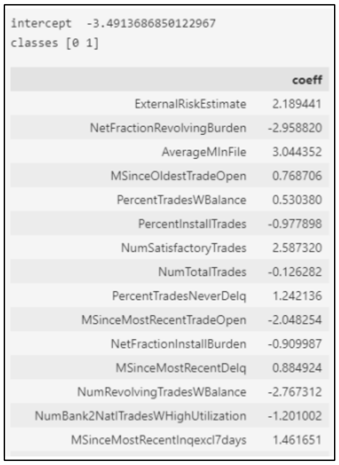


Fig 5. SHAP

Logistic Regression Model:

The third model used for the website was a logistic regression model. This model was used for a similar reason as the simple decision tree, it is fairly accurate, and it is fairly easy to understand how it came to its prediction. The logistic regression model works by forming a gradient coefficient for each feature used in the model then combines them all into a long equation for each feature of each point, then if the equation is greater than one it makes one prediction and if it is below 0 it makes a different prediction. Therefore, this equation can be fairly understood by looking at each coefficient for each of the features with bigger coefficients having a greater effect on the outcome of the prediction. A table of the coefficients given to each of the features in our dataset can be seen in figure 6 seen to the side.



```
intercept  -3.4913686850122967
classes [0 1]
```

	coeff
ExternalRiskEstimate	2.189441
NetFractionRevolvingBurden	-2.958820
AverageMInFile	3.044352
MSinceOldestTradeOpen	0.768706
PercentTradesWBalance	0.530380
PercentInstallTrades	-0.977898
NumSatisfactoryTrades	2.587320
NumTotalTrades	-0.126282
PercentTradesNeverDelq	1.242136
MSinceMostRecentTradeOpen	-2.048254
NetFractionInstallBurden	-0.909987
MSinceMostRecentDelq	0.884924
NumRevolvingTradesWBalance	-2.767312
NumBank2NatlTradesWHighUtilization	-1.201002
MSinceMostRecentInqexcl7days	1.461651

Fig 6. Logistic regression coefficients

Justification of Model Choice:

Below are all the models that were made with their corresponding accuracies, however, only the first 3 are chosen on the website for several reasons outlined below. When comparing and testing the models, a 5-fold cross validation was performed on the dataset for the model and the average of the 5 folds was taken to be the accuracy score of the models. The scoring metric used for the testing is ROC AUC. This is because the models are performing a binary classification and AUC measures the model's ability to correctly distinguish between the 2 different positive and negative classifications and return this as an easy-to-understand summary.[1] ROC AUC metric is also best used when there is a good balance of each class in the dataset which for the data used there is, as can be seen below in figure 7.

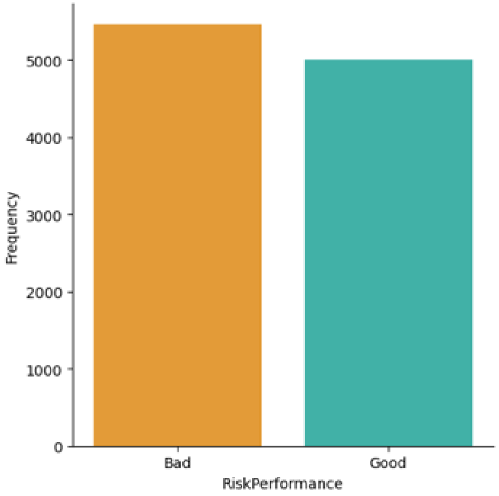


Fig 7. Class Balance

Table 1 All models accuracy comparison

Model	Accuracy (ROC AUC, 5-fold cross validation average)
Decision Tree	0.74
LightGBM	0.79
Logistic Regression	0.79
Naïve Bayes (NB)	0.76
KNN	0.78
Neural Network	0.69

As can be seen in table 1 above, the LightGBM and the logistic regression model had the highest accuracy. Fortunately, they are also explainable models, and it is easy to see how each model made its decision using LIME and SHAP python libraries. The third model used was also the first most simple model produced, the simple decision tree classifier. This model was also used because it is the most easily explained model and the clearest model in setting out each step of its decision-making process and why it came to its conclusion. Therefore, despite being less accurate than the KNN and NB models, it is important to have the easy-to-understand explainability for the sake of clients that may be rejected for loans and credit by a prediction made by our website. The other two models added are also both more accurate than KNN and NB and will also have explainability built into them and shown on the website. As this is a minimum viable product, it is envisioned in future updates that these three used models will be all put together into an ensemble and will each be able to be used at once to generate a single risk prediction with a single set of explanations, rather than needing to run all models separately and to compare that the prediction from all three all matches.

As can be seen in table 1, a more advanced neural network machine learning network was also produced to see if using a more advanced model will produce a more accurate prediction than the simpler models. However, it was found that these models make it very hard to see how they made each decision, consequently they can be known as a "black box" algorithm due to the inner workings of the algorithm and how it gets to its prediction being a mystery. Therefore, after initial optimisation of the algorithm, the accuracy of the neural network was less good than the more basic models, so it was decided a more advanced neural network was not needed as explainability was just as important as accuracy to ensure fairness of the website when used in a real-life setting.

3.4 Front End

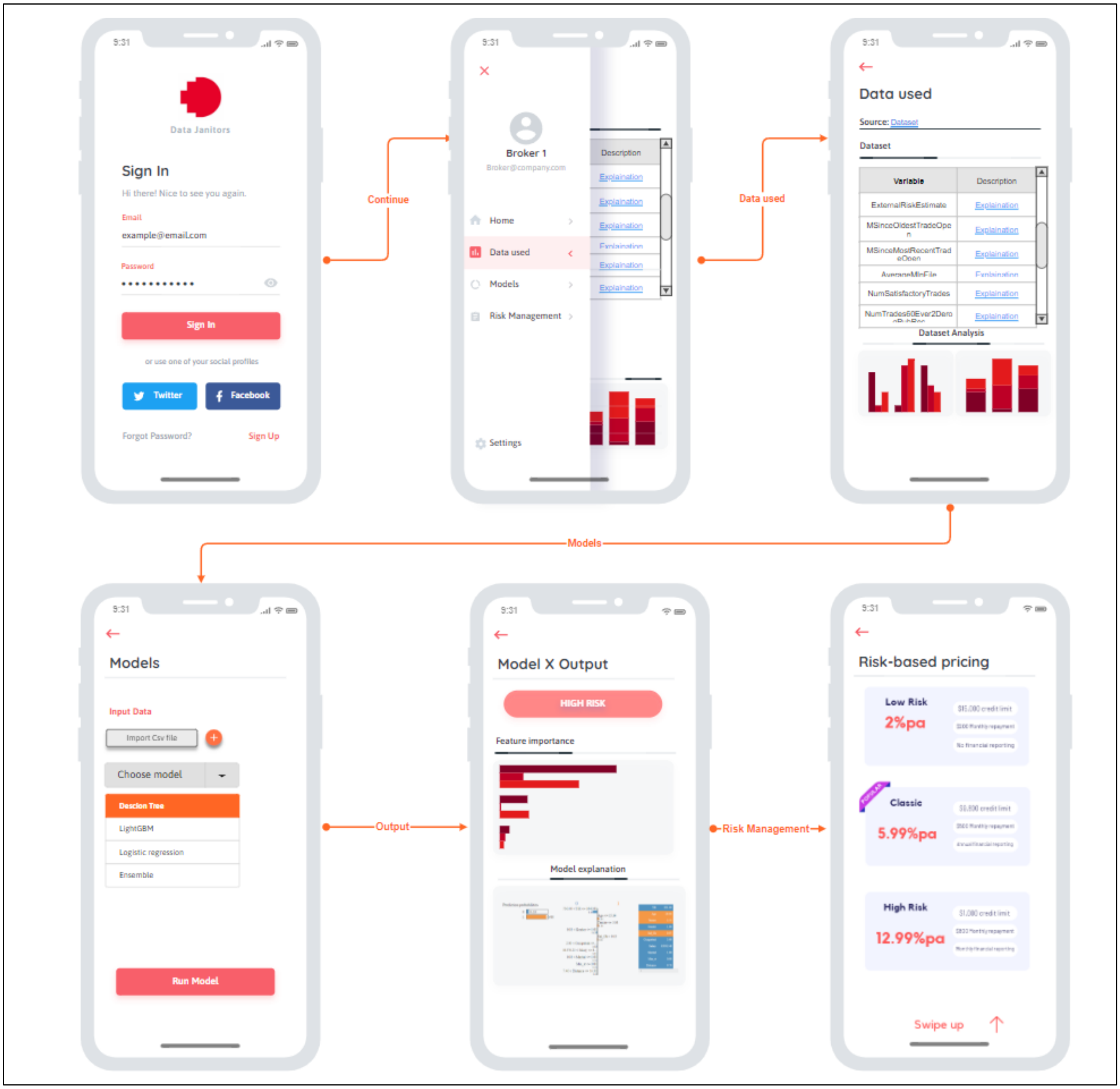
As one of the main objectives of the front end/application is to provide small brokers with an explainable easy to use loan risk prediction software, the main requirements that we considered were as follows:

1. The application **MUST** use data that is relatively straightforward and easy to obtain for financial institutions
2. The application **MUST** not contain unethical bias towards marginalized groups as such a model would contradict the equitable aims of our application
3. Each model must be explainable (No black boxes)

4. Each decision/output made by any model must be explained in a concise manner with limited jargon (e.g., showing hundreds of decision trees for a single boosted regression model is not feasible)

UI Mock-up

To get a basic idea of how our UI will look and what front end software would be feasible to generate a desirable minimum viable product we desired to first mock-up the ideal fundamental design elements in app form as seen below.



Variable	Description
ExternalRiskEstimate	Explaination
MSinceOldestTradeOpen	Explaination
MSinceMostRecentTradeOpen	Explaination
AverageMInFile	Explaination
NumSatisfactoryTrades	Explaination
NumTrades60Ever2DerogPubRec	Explaination

Fig 8. UI mock-up

For aid in programming and to just generally consolidate the functionality requirements of our design we also created a basic flow chart as to how our front end would function as seen below.

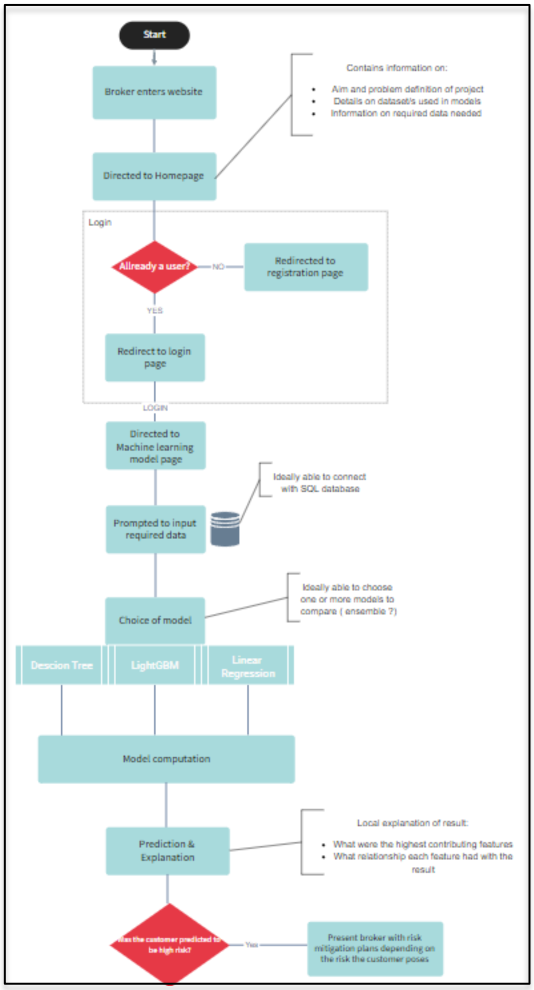

Start
Broker enters website
Contains information on:
• Aim and problem definition of project
• Details on dataset/s used in models
• Information on required data needed
Directed to Homepage
Login
Allready a user?
NO
Redirected to registration page
YES
Redirect to login page
LOGIN
Directed to Machine learning model page
Ideally able to connect with SQL database
Prompted to input required data
Ideally able to choose one or more models to compare (ensemble ?)
Choice of model
Descion Tree
LightGBM
Linear Regression
Model computation
Local explanation of result:
• What were the highest contributing features
• What relationship each feature had with the result
Prediction & Explanation
Was the customer predicted to be high risk?
Yes
Present broker with risk mitigation plans depending on the risk the customer poses

Fig 9. Flow chart of design

Final Product

Due to time constraints and the difficulty in app development as well as the simple fact that most users of the product would require it only once, we saw fit to do away with the idea of an app and steam ahead with a web-based client product. The back end of the final website was written in python Flask, due to its high flexibility, ease of use and its compatibility with our machine learning models which were already coded in python. As the models themselves don't require a large amount of storage space we were also able to host our flask application for free using Heroku's cloud platform which was compatible with flask.

LINK: https://datajanitors.herokuapp.com/

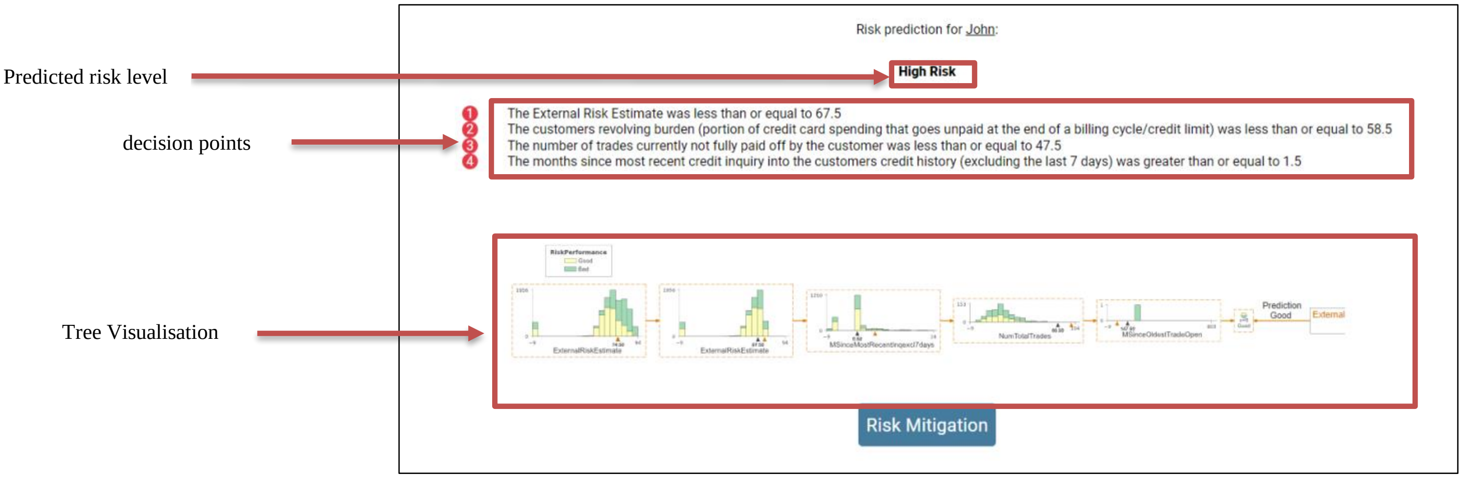
INPUT FOR EACH MODEL: CSV file containing customer name and the values of all 23 features of the dataset. e.g.:

Customer name	MSinceOldestTradeOpen	MSinceMostRecentTradeOpen		NetFractionInstallBurden	PercentTradesWBalance
Matt	55	144	…	1	69

Decision Tree UI

The decision tree model being the most straightforward in terms of explainability required a relatively simplistic output and is probably the easiest model for a broker to understand. The final output for each user inputted in the csv consists of:

1. The Predicted risk level of the user (in terms of their ability-to-repay credit): High risk or Low risk
2. A list of all the decision points considered and their outcome for the given user
3. Visualization of the overall decision pathway used to make the decision



Risk prediction for John:

High Risk

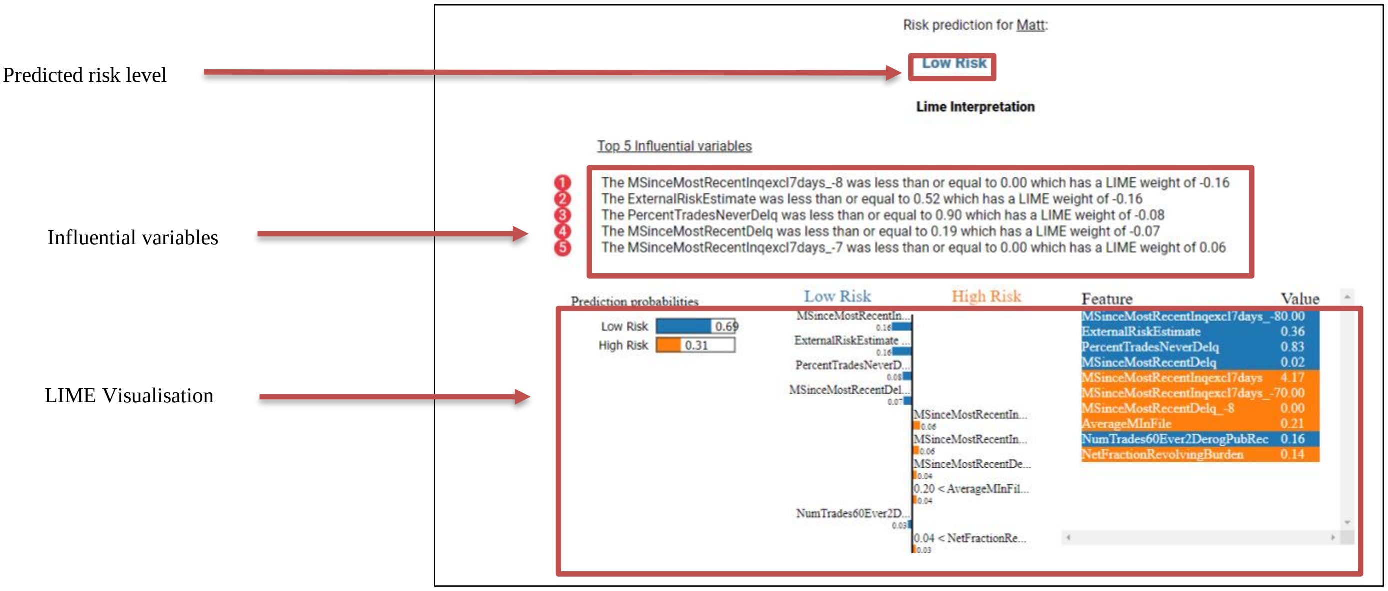
1. The External Risk Estimate was less than or equal to 67.5
2. The customers revolving burden (portion of credit card spending that goes unpaid at the end of a billing cycle/credit limit) was less than or equal to 58.5
3. The number of trades currently not fully paid off by the customer was less than or equal to 47.5
4. The months since most recent credit inquiry into the customers credit history (excluding the last 7 days) was greater than or equal to 1.5

Fig 10. Decision Tree implementation

LightGBM UI

Because our LightGBM model consisted of 200 weighted decision trees, using the same methods used for the decision tree model would not be feasible for an everyday user to interpret for every customer tested. To reduce this complexity and facilitate ease of use we decided to use LIME which uses linear models to approximate the local behaviour of our LightGBM model and then subsequently ranks the features based their overall contribution to the models. The final output for each user inputted in the csv consists of:

1. The Predicted risk level of the user (in terms of their ability-to-repay credit): High risk or Low risk
2. The top 5 most influential variables and their respective LIME weights
3. Visualization of the overall LIME output containing prediction probabilities, LIME weights and overall contribution of each variable



Risk prediction for Matt:

Low Risk

Lime Interpretation

Top 5 Influential variables

1. The MSinceMostRecentInqexcl7days_-8 was less than or equal to 0.00 which has a LIME weight of -0.16
2. The ExternalRiskEstimate was less than or equal to 0.52 which has a LIME weight of -0.16
3. The PercentTradesNeverDelq was less than or equal to 0.90 which has a LIME weight of -0.08
4. The MSinceMostRecentDelq was less than or equal to 0.19 which has a LIME weight of -0.07
5. The MSinceMostRecentInqexcl7days_-7 was less than or equal to 0.00 which has a LIME weight of 0.06

Prediction probabilities
Low Risk 0.69
High Risk 0.31
Low Risk
High Risk
MSinceMostRecentIn... 0.16
ExternalRiskEstimate ... 0.16
PercentTradesNeverD... 0.08
MSinceMostRecentDel... 0.07
MSinceMostRecentIn... 0.06
MSinceMostRecentIn... 0.06
MSinceMostRecentDe... 0.04
0.20 < AverageMInFil... 0.04
NumTrades60Ever2D... 0.03
0.04 < NetFractionRe... 0.03

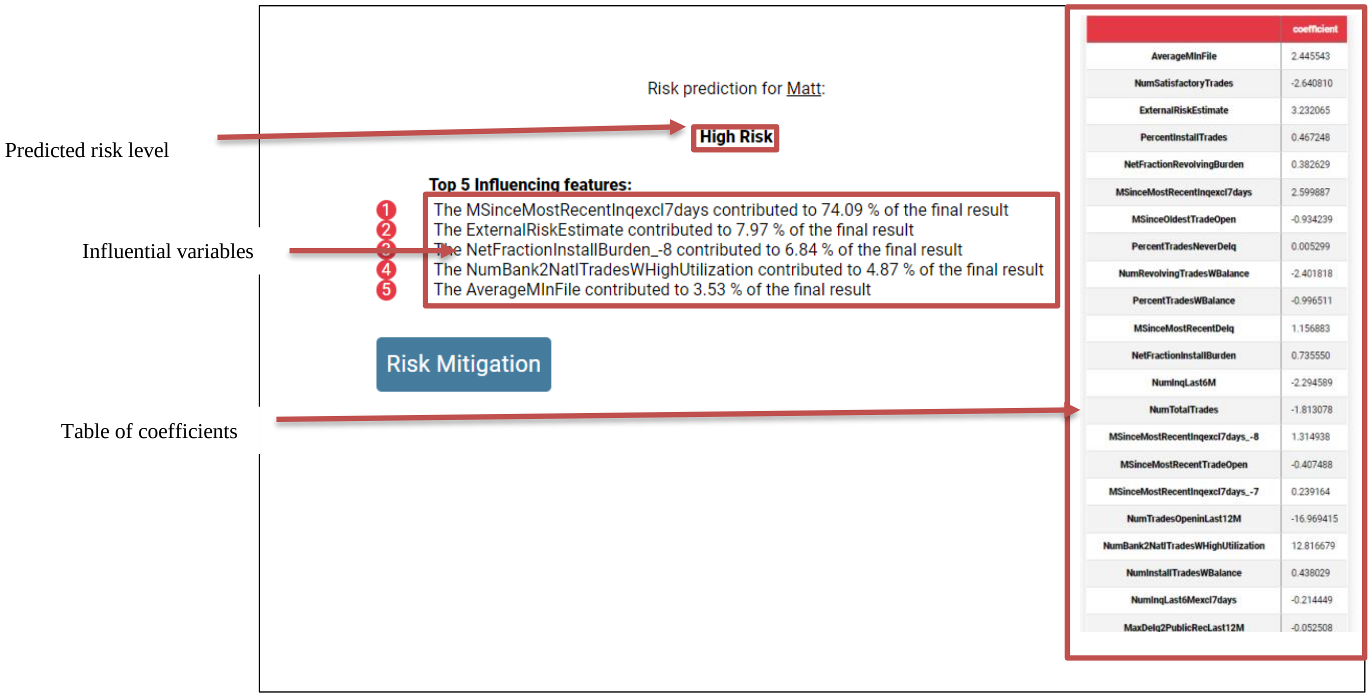
Feature	Value
MSinceMostRecentInqexcl7days_-80.00	
ExternalRiskEstimate	0.36
PercentTradesNeverDelq	0.83
MSinceMostRecentDelq	0.02
MSinceMostRecentInqexcl7days	4.17
MSinceMostRecentInqexcl7days_-70.00	
MSinceMostRecentDelq_-8	0.00
AverageMInFile	0.21
NumTrades60Ever2DerogPubRec	0.16
NetFractionRevolvingBurden	0.14

Fig 11. LightGBM implementation

Logistic Regression UI

Like the decision tree model, logistic regression is relatively straightforward in its methodology and explainability, however the model itself is likely not as intuitive as the decision tree for everyday users therefore further processing of the output needs to be performed. To generate a more palatable output, instead of just presenting the coefficients and intercept of the final model we decided to compute the contribution that each variable has on the result with respect to the other variables (not overall contribution) using the models' coefficients, intercept, and input user values. The final output for each user inputted in the csv consists of:

1. The Predicted risk level of the user (in terms of their ability-to-repay credit): High risk or Low risk
2. The top 5 contributing features to the predicted risk level based off coefficients
3. Table of coefficients for the logistic regression model



Risk prediction for Matt:

High Risk

Top 5 Influencing features:

1. The MSinceMostRecentInqexcl7days contributed to 74.09 % of the final result
2. The ExternalRiskEstimate contributed to 7.97 % of the final result
3. The NetFractionInstallBurden_-8 contributed to 6.84 % of the final result
4. The NumBank2NatlTradesWHighUtilization contributed to 4.87 % of the final result
5. The AverageMInFile contributed to 3.53 % of the final result

	coefficient
AverageMInFile	2.445543
NumSatisfactoryTrades	-2.640810
ExternalRiskEstimate	3.232065
PercentInstallTrades	0.467248
NetFractionRevolvingBurden	0.382629
MSinceMostRecentInqexcl7days	2.599887
MSinceOldestTradeOpen	-0.934239
PercentTradesNeverDelq	0.005299
NumRevolvingTradesWBalance	-2.401818
PercentTradesWBalance	-0.996511
MSinceMostRecentDelq	1.156883
NetFractionInstallBurden	0.735550
NumInqLast6M	-2.294589
NumTotalTrades	-1.813078
MSinceMostRecentInqexcl7days_-8	1.314938
MSinceMostRecentTradeOpen	-0.407488
MSinceMostRecentInqexcl7days_-7	0.239164
NumTradesOpeninLast12M	-16.969415
NumBank2NatlTradesWHighUtilization	12.816679
NumInstallTradesWBalance	0.438029
NumInqLast6Mexcl7days	-0.214449
MaxDelq2PublicRecLast12M	-0.052508

Fig 12. Logistic regression implementation

Risk Management

In cases where high risk is predicted, the user is directed towards our risk mitigation page which displays our current three proposed risk mitigation plans that can be deployed/offered to customers with high risk to reduce the risk they pose to the business without outright denying a potential customer service.

Our proposed mitigation plans are as follows:

1. **Risk based pricing:** Risk base pricing is where a lender can tailor their interest rates based on the predicted risk value, meaning higher risk clients are given higher interests rates/credit costs to provide a buffer for potential cash flow disruption and cost of recovering loans. Additionally, lenders can send lower credit limits on the amount that can be borrowed to reduce the overall burden that the risk poses. [12]

2. **Financial covenants:** Financial covenants refers to a technique for managing high risk clients by forming a debt covenant with said client. A debt covenant is when a lender establishes a predetermined agreement with a client that sets specific regulations/conditions that the borrower must meet to gain access to the specified credit. This covenant can also include a higher set leverage ratio, meaning the lenders require a higher proportion of collateral from the client when compared to the borrowed credit. [5]

3. **Financial reporting:** Occurs when high risk clients are, to maintain a credit agreement, required to send financial statements and/or tax records to lenders for proof of income and expenses. This allows lenders to accurately identify a borrower’s ability-to-repay and therefore is more capable of taking the required steps to recover funds before repaying becomes too hard for the lender. Depending on the risk level of the client these reports can be required monthly, quarterly, semi-annually, or annually.

4. Findings

Step	Timeframe
Pre-approval	1-3 days
Application	3-5 business days
Property valuation	3-5 business days
LMI (if it applies)	1-2 days
Loan approval & settlement	4-6 weeks

Table 2: Approval Timeline

As per our research [9] the average time for an entire process of a home loan to be approved and a settlement to be generated is around 5-8 weeks depending on the various conditions.

- Pre-approval phase would be to understand all the various types of loans available and how much loan can be granted based on the property in question.(1-3 days)
- Application phase would be filling in all the necessary paper works and collected all the necessary documents required to be sent along with the application.(3-5 business days)
- Property valuation phase would be the estimated price of the selected property will be arranged by the financial institution and lenders. (3-5 business days)
- LMI is the Lenders Mortgage Insurance here if the more than 80% of the value of the property is borrowed. (1-2 days)
- Loan approval & settlement: This is final phase after all the approvals are achieved from the previous stages and the credit checks are positive the loan will be granted.

Overall, the process of applying a loan with a bank can vary from 5-8 weeks depending on various conditions. The entire process can be quite tedious and time consuming as well as most of the time the user has no idea why exactly their loan has been rejected.

We aim with our project to help all the stakeholders in a loan application process, by providing a way in which will benefit both the individual user and financial institution by providing them a way to check if the respective user is a high-risk client or not also provide a reason why this is the case. Thus, saving on a significant amount of time in going through the application process even though the credit profile is not good enough.

5. Impact and Significance of Results

The central goal of our project was to be able to help empower people and organizations who were less privileged and help them take equitable and accountable data driven decisions. There are already models being used by all the big banks and financial institutions which are trained by highly skilled data scientists, which most likely outperform ours. We as a team and aspiring data scientists, have a core belief that the direction that you are going in and the value it provides to the real world, matters much more than just trying to use the fanciest tech trying to make the best model possible with the highest prediction accuracies, all for nothing.

Another thing is, since our website was made keeping in mind, the people, rather than just being made for our personal gain and trying to make the most money out of it as possible, it promotes more fair and unbiased results which of course can be cross verified with the all the transparency we provide. It's quite likely with the banks that, even though they might find 2 candidates that are both fit to repay a loan, they might choose the one who is more likely to keep making them money in the future, because it's a business, with a profit-centric approach, and not people-centric. Hence, we as a team are aiming to ensure that we stay focused on the people, the tools that provide them with the most value and the tools to understand the workings behind the scenes.

5.1 Impact on lenders/ financial institutions

Since our website is open source, it will allow lenders and small financial institutions to use a more grounded and robust way to find out if a candidate is fit to be able to make a repayment or not. Currently, the lenders generally don't have access to very proficient software that is able to ascertain these specific answers. The current software being utilised is mostly hard-coded or employs credit scores, which only cover your past loan history, rather than your current capabilities. So, if a person was not capable or able to make some repayments on time 20 years ago, the current systems don't consider the changes the specific person has made over that period, the user is then penalised and is left at a significant disadvantage.

Also, in more backward regions and in developing countries, lenders have been depending purely on the word of mouth and references to know whom to risk and whom to not. Which of course is not very precise, but secondly, has more chances to sneak in unfairness and personal biases, such as discrimination based on race, caste, sex, colour, etc. Our model aims to bring an end to this and provide everyone regardless of economic status and societal factors the tools necessary to build their lives and increase their living standards.

5.2 Impact on individuals

Credit score is a number assigned to a credit profile of an individual from a range of 300 to 850 based on the creditworthiness of that individual[7]. The higher the score, the better the chances of a loan approval.

This number is generated based on the previous credit history of the customer. Factors that affect credit score involve the individual’s payment history, total amount owed by the individual, length of their credit history, types of credit (i.e., types of loans, credit cards, or instalment credits) and the number of new buy-now-pay-later accounts owned by the individual.

Having a bad credit score will lead to the credit request put forth by an individual to be rejected. Some reasons for a bad credit score[2] might be that the individual has multiple loans with various lenders, they have defaulted on an Emi on a current loan, there is a history of defaulted payments in their payment history, etc. All these factors have a negative impact on the credit score and leave them at a disadvantage.

Our website is a double-edged sword, so it helps lenders to get more accurate predictions as well as helps the lenders give the individuals, their right to information and help them understand why their loan got rejected. The info that they get, paired with the risk mitigation strategies that we offer, will help people to know what they could work on to make their profile better, and increase their chances of getting the loan accepted, rather than just themselves having to go through tons of resources to find that out, or having to do some financial consulting. This will of course also help the lenders get in more capable customers with strong profiles. So, the explainability gives a win-win situation to both parties.

The other benefit that our website will provide to individuals is that it being able to be accessed by anyone, both individuals and money lenders/institutions, will save the individuals from the hassle of going through the lengthy process of loan applications and then waiting for the outcome of their loan applications for over 6 weeks, just to get turned down by the bank. They can just hop on our website and have a rough estimate if they are anywhere near of getting their loan accepted. Of course, the actual outcome of the bank may vary because there are different systems in place, but they will still get a pretty good idea. In the future, if we get enough support, we might even make the whole process automated and generate an accredited certificate which individuals can generate for themselves and use it directly with lenders to show their repayment capabilities and will make the loan applications very fast. It could also be used with their bank loan applications to strengthen their profile further.

5.3 Value of Project

The value of the project is immense as it is built to help all the stakeholders involved in the loan application process. Using this website individuals and lenders can deem whether their profile or the profile given to them is fit at the moment to be used to apply for loans with a financial institution or not. If it is not, they can follow the mitigation plans to reduce the credit risk in a profile before applying to a loan process. Thus, saving on considerable time and effort of the stakeholders in identifying whether the loans will be approved or not.

Even helping the candidate [8] with a bad credit profile by suggesting them not to apply for a loan to keep a clean credit report without any previous inquiry records present which will help in applying for a loan the next time. Being an open-sourced project also means that the inner workings can be seen, these ideas and processes such as increased transparency and explainability can then be used to improve upon current processes in the financial industry as a whole.

5.4 Business case

The initial problem is that in the current environment there are not many ways for a customer to understand why their loan was not approved and what contributed to their respective to loan application to be rejected.

This leads to a lot of back-and-forth communication between the customers and the lenders impacting significantly on the time of both customers and lenders in the process. Our project aims to reduce this impact on time by providing a means for the customer or lender to go through the credit risk of a profile before creating an application and understand the reasons for the risk of the profile.

This will benefit the individuals by letting them work on their credit profiles as they will better understand why their profile is categorized as elevated risk and the financial institution by not getting profiles with high-risk credit in the loan application process. Thus, saving time for all the parties involved in the process. Also, saving the loan candidate from a loan rejection appearing in their credit report which may lead to additional inquiries when applying the next loan.

In the future, revenue can be derived by charging small single use access fee or a subscription-based model for small-scale lenders. Revenue from single use private users will be derived from an ad-based model. These two revenue streams together could possibly provide, if implemented well, a strong source of revenue and subsequently some profits to ensure the upkeep of the website and the refinement of the model.

6. Future Improvements

Due to time constraints and scope of the overall project, we obviously couldn’t produce the perfect ready to launch application, however as the main objective of producing a minimal viable product is to take a more iterative approach to product design, identifying areas for possible improvements and their feasibility is just as important as the project itself.

Some suggestions for improving upon the current application are:

- Allow the application to generate a single output for each input which would contain:
  - An Ensemble model which utilizes the predictions computed by all three models and generates weighted explanation based off a summary of all local explanations by all three models.
  - The results and explanations for each model used respectively
  - A summary table showing the current scores for each model (including the ensemble) based on cross validation results with a large test set.

This allows for borrows to not only make a more informed decision on which model to use, but also saves time in having to input a csv file for every model to compare results.

- Generate more user-specific risk mitigation methods and maybe even calculate the optimal risk-based pricing, reporting period etc. based on the level risk that a customer holds.
- Further exploration in the use of additional datapoints and its effect on accuracy without introducing additional bias.
- Bias mitigation and fairness implementation through the use of AI Fairness 360 or Google's What-If
- Integrate the app with a general borrows database, to allow brokers to easily implement such a system in the workplace which can consistently be updated with new information regularly and easily.
- The website also needs to be adequately tested for any potential undetected bias in the data used to train the models to ensure they don't treat a particular feature too harshly when predicting risk. Such as treating a person who has never had a loan or has had bad debts a long time ago an unfair prediction.

7. Project Management

7.1 Task Breakdown and clear outline

To optimize the entire workflow of the project and to make our plans/ideas more manageable, especially with most team members having a busy work schedule outside the project, we decided to utilize a more compartmentalized and lean approach. We achieved this by first segregating the 4 main overarching tasks of the project into teams, these consisted of:

1. Front end development
2. Back-end development
3. Report writing
4. Presentation

Obviously, these teams remained relatively fluid as tasks like report writing and the presentation was not needed till the second half of the project, and everyone was still able to contribute to some aspect of every team. However, having a distinct team leader for each task allowed for a much more efficient workflow and higher level of organisation. Additionally, as our goal was to produce a minimal viable product efficiently, we decided to adapt the lean methodology and take a more iterative approach. To apply this for both the back and front-end development we decided to take it in phases so that we could regularly assess our situation after each phase and identify the best course of action from that point.

Front end development phases:

I. Dataset analysis and pre-processing
II. Development and evaluation of machine learning models (#1)
III. Explore explainability options of models (#1)
IV. Development and evaluation of machine learning models (#2)
V. Explore explainability options of models (#2)
VI. Finalise best models for first MVP

Backend end development phases:

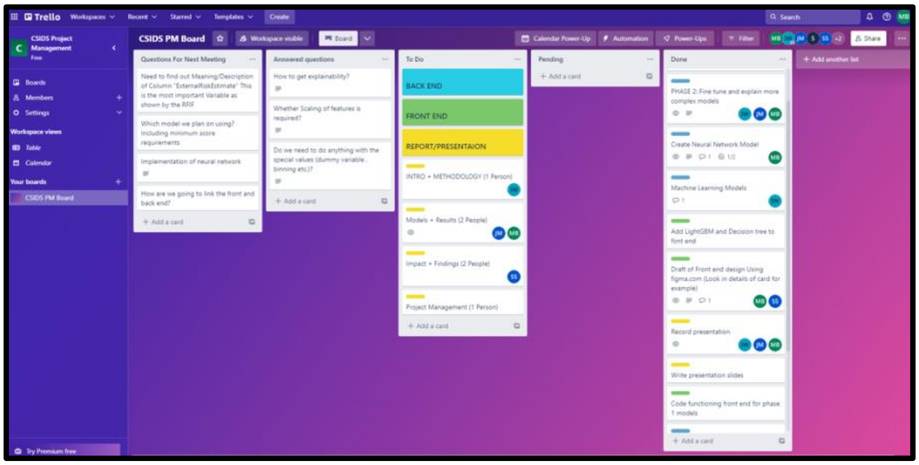
I. Develop plan of attack (Flow charts & UI mock-ups)
II. Research front end approaches
III. Test model functionality on desirable front end approach
IV. Add basic explainability functionality
V. Test on addition models and improve upon explainability
VI. Finalise font-end MVP

7.2 Collaboration

We used Trello boards on a regular basis to maintain a high level of organisation during various levels of our product development stages. We found Trello to be a very useful tool in situations where not everyone was available to meet or had conflicting schedules, as it enabled contact communication and progress reporting outside of our meetings. The key benefits of using a project management board like Trello was that:

- It allowed us to keep track of questions for each other and for Khaled
- Set expectations and roles for all group members throughout the project
- Broke up the project into more manageable stages
- It allowed for the front end and backend team to keep tabs on each other's progress

For the analysis we used Github in conjunction with VS code to jointly develop the code required for the front and back end.



Questions For Next Meeting

- Need to find out Meaning/Description of Column "ExternalRiskEstimate" This is the most important Variable as shown by the RFIF
- Which model we plan on using? Including minimum score requirements
- Implementation of neural network
- How are we going to link the front and back end?

Answered questions

- How to get explainability?
- Whether Scaling of features is required?
- Do we need to do anything with the special values (dummy variable , binning etc.)?

To Do

- BACK END
- FRONT END
- REPORT/PRESENTAION
- INTRO + METHODOLOGY (1 Person)
- Models + Results (2 People)
- Impact + Findings (2 People)
- Project Management (1 Person)

Pending

Done

- PHASE 2: Fine tune and explain more complex models
- Create Neural Network Model
- Machine Learning Models
- Add LightGBM and Decision tree to font end
- Draft of Front and design Using figma.com (Look in details of card for example)
- Record presentation
- Write presentation slides
- Code functioning front end for phase 1 models

Fig 13. Trello board

7.3 Weekly meets and checkpoints

Weekly meets and updates were a priority to ensure all the members are on the same page and working towards a common goal. Checkpoints helped us to stay on course with the deadlines ensuring we have completed the required parts as and when required.

7.4 Gantt Chart

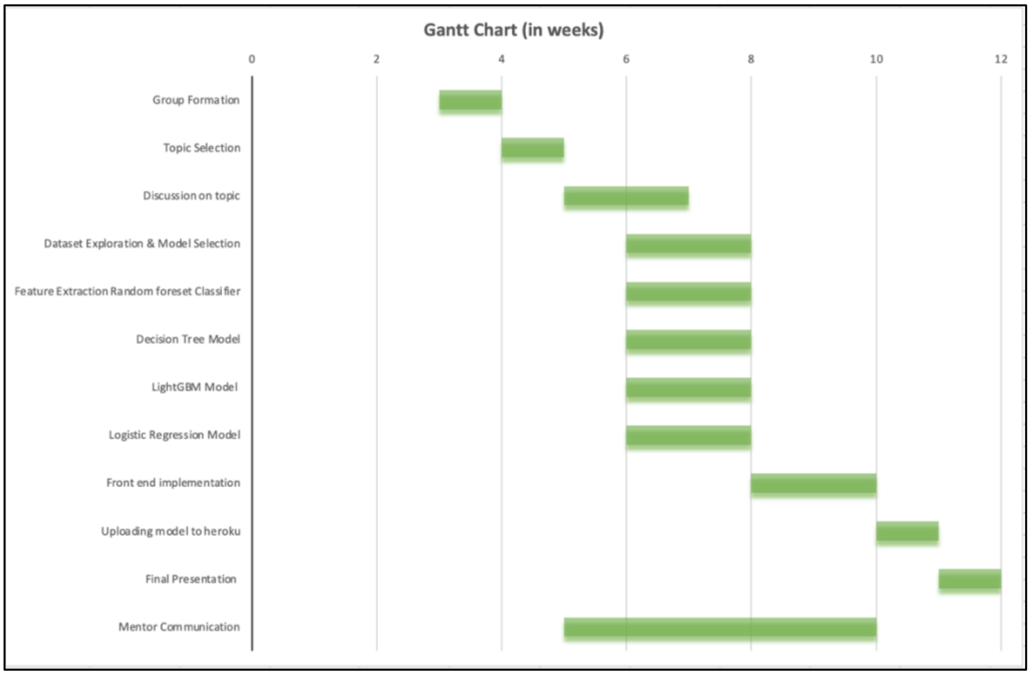


Fig 14. Gantt chart

7.5 Contributions

Below is a table outlining the individual contributions and roles each team member provided to the project

Team Member	Contributions
Matthew Bentham	• Front end planning and development • Report writing (Front end & project management sections) • Backend model testing and identifying explainability methods
John Fergus Murrowood	• Front end development and risk mitigation strategy research • Development of LightGBM model • Report sections (Problem statement, backend methodology, future improvements)
Alex Peter Thomas	• Report writing (Structure, Introduction, Impact and Significance, Findings, sections in Methodology, Project Management) • Dataset Selection & Exploration • Frontend & Backend discussions and planning
Isxaq Warsame	• Backend Model Development, front end planning • Report writing (Planning, sections in Into, Project Management) • Presentation • Final Editing
Udit Dinesh	• Report writing (Structure, sections in Methodology, Project Management) • Dataset Selection & Exploration • Frontend & Backend discussions and planning
Simran Sidhanti	• Report writing (Structure, sections in Methodology, Project Management) • Frontend planning and development • Frontend & Backend discussions

8. Special Thank You

Special thanks to our mentor & supervisor Khaled Iqbal for taking the time to meet with us weekly to help guide the direction of the project and answer any clarifying questions we had. These meetings helped us keep on track and aided us in developing and altering our weekly expectations/tasks to best meet our problem statement. He was instrumental in our implementation and provided much needed guidance and advice while still allowing for our autonomy. Big thank you for the advice and kudos to Khalid for being a vital instrument in our final achievements

9. References

[1] A. Bhandari, “AUC-Roc Curve in machine learning clearly explained,” *Analytics Vidhya*, 14-Jun-2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/. [Accessed: 21-Oct-2022].

[2] Bank Bazaar, “6 reasons why your credit card application can be rejected,” Compare & Apply for Credit Cards & Loan Online in India, 2022. [Online]. Available: https://www.bankbazaar.com/credit-card/reasons-why-your-credit-card-application-can-be-rejected.html. [Accessed: 21-Oct-2022].

[3] C. Bento, "Decision Tree Classifier explained in real-life: picking a vacation destination", *Medium*, 2021. [Online]. Available: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575. [Accessed: 28- Sep- 2022].

[4] Fico community, “Explainable Machine Learning Challenge”, Fico Community, 2018. [Online]. Available: https://community.fico.com/s/explainable-machine-learning-challenge. [Accessed: 28- Sep- 2022].

[5] “Financial Covenants - Overview, Importance, Examples.” https://corporatefinanceinstitute.com/resources/knowledge/finance/financial-covenants/ [Accessed 16-Oct- 2022].

[6] “Gradient Boosting Machines,” *Gradient Boosting Machines · UC Business Analytics R Programming Guide*. [Online]. Available: http://uc-r.github.io/gbm_regression. [Accessed: 21-Oct-2022].

[7] Investopedia, “Credit score: Definition, factors, and improving it,” Investopedia, 26-Sep-2022. [Online]. Available: https://www.investopedia.com/terms/c/credit_score.asp. [Accessed: 20-Oct-2022].

[8] J. Ulzheimer, “Does a declined loan appear on Your credit report?” Experian, 11-Aug-2020. [Online]. Available: https://www.experian.com/blogs/ask-experian/does-a-declined-loan-appear-on-your-credit-report/. [Accessed: 21-Oct-2022].

[9] loans.com, "How long does it take to get a home loan approval?", Loans.com.au, 2022. [Online]. Available: https://www.loans.com.au/home-loans/how-long-does-it-take-to-get-a-home-loan-approval#:~:text=The%20average%20time%20for%20formal,reaching%20settlement%20on%20the%20property. [Accessed: 18- Oct- 2022].

[10] Nitin, "LightGBM Binary Classification, Multi-Class Classification, Regression using Python", *Medium*, 2020. [Online]. Available: https://nitin9809.medium.com/lightgbm-binary-classification-multi-class-classification-regression-using-python-4f22032b36a2. [Accessed: 28- Sep- 2022].

[11] ODSC, "The Importance of Explainable AI", Medium, 2018. [Online]. Available: https://medium.com/predict/the-importance-of-explainable-ai-28db06e0c802. [Accessed: 18- Oct- 2022].

[12] W. Edelberg, “Risk-based pricing of interest rates for consumer loans,” *J Monet Econ*, vol. 53, no. 8, pp. 2283–2298, Nov. 2006, doi: 10.1016/J.JMONECO.2005.09.001. [Accessed: 16-Oct-2022].