Text

Description automatically generated

### Loan Origination with Machine Learning and Robotic Process Automation

### Introduction

When a bank embarks on a journey to approve a loan or a mortgage, it needs to take a number of decisions. The end result needs to be a loan or a mortgage that is profitable to the bank, resulting in a satisfied customer who thinks they got a fair interest rate, with minimal risk of loan default for the bank.

In this article, I will talk about how Loan Origination Systems can use Machine Learning and Robotic process automation to reduce Loan processing time significantly while reducing the risk of default to the bank. All this results in a satisfied customer.

I will go through the following sections in this article.

1. About Loans and Mortgages and the Loan Process
2. Using RPA in Application Processing
3. Machine Learning models that Loan Origination would need
4. Steps to create an ML model. Special call-out for Step 1 where we talk about the input data for our models
5. Options for small lenders who don't have access to BigData

### Loans and Mortgage Process

Let us start by discussing the Loans and Mortgage process and identify the bottlenecks.

Graphical user interface, diagram, text

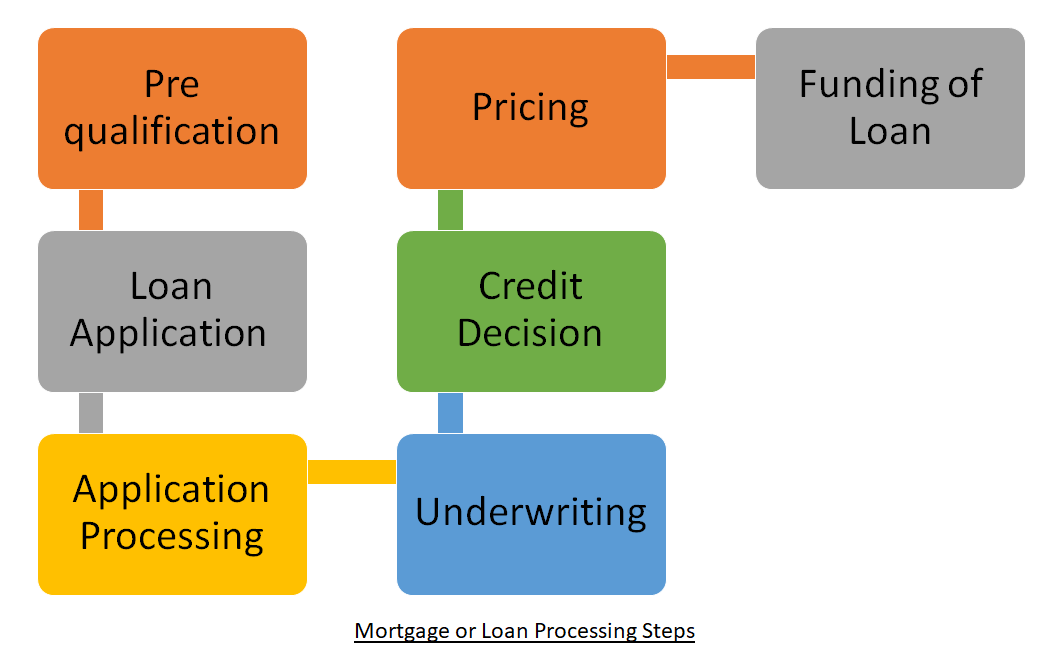
Description automatically generated

There are two types of loans

1. Secured loans: called Mortgages. They are used for buying property. Since the loans are secured, the banks can offer them at a lower interest rate. Banks can seize the property if the loan is not paid back, hence the word secured.

2. Non secured loans: called Loans. They are used for buying anything that isn’t attached as collateral e.g. a loan for a vacation. They are generally offered at a higher interest rate compared to secured loans.

The lenders for the loan, say banks or building societies, use a Loan origination system (LOS) to process the loan application. Processing a loan typically has the following steps.



1. Pre-qualification — borrower collects all documents required and submits them to the lender who may choose to pre-qualify the loan allowing the borrower to proceed to the next steps.

2. Loan Application — borrower submits online or paper application.

3. Application Processing — lender reviews application for accuracy and completeness.

4. Underwriting — lender performs credit and risk evaluation.

5. Credit Decision — Generally one of three decisions are made: approved, denied, or approval with modified loan parameters.

6. Pricing — price the loan by determining the interest rate, duration, and installment amount. This is generally the secret sauce of each lender.

7. Funding of Loan — sign paperwork and disburse the funds to the borrower.

The above steps combined take an average of 30 to 45 days. This can come down significantly with RPA automation and machine learning. The loan default risk of the bank can also be managed better using machine learning.

### Robotic Process Automation for Application Processing

Let us first define Robotics Process Automation or RPA.

According to a Mckinsey report [1], RPA is a type of software that mimics the activity of a human being in carrying out a task within a process. It can do repetitive stuff more quickly, accurately, and tirelessly than humans, freeing them to do other tasks requiring human strengths such as emotional intelligence, reasoning, judgment, and interaction with the customer.

Chart, bubble chart

Description automatically generated

One of the time-consuming steps in Loan processing is “Application processing”. This step can be speeded up by Robotic Process Automation. The manual processing of all the documents, reports, and forms can be error-prone, and also takes time. There are additional manual steps involved when an error or missing information is identified, where the loan officer will need to get back to the borrower to get the missing information.

From an application processing point of view, we can think of RPA as software that can be configured to read all the documents, online portals, and extract fields out of them to populate the lender’s Loan Origination system. It can be used to read documents, identify missing information, send an email to the borrower for additional information, and process the email response when received. This can make the Application process many times faster.

### Machine Learning for Loan Underwriting

The banks, who are generally the lenders, are sitting on a gold mine of data. They also have access to government data, like historical pricing and future area development plans. Using this Big Data, lenders can create a much better prediction of risk and hence offer more competitive or conservative interest rates.

#### Machine Learning Models needed for Loan Origination

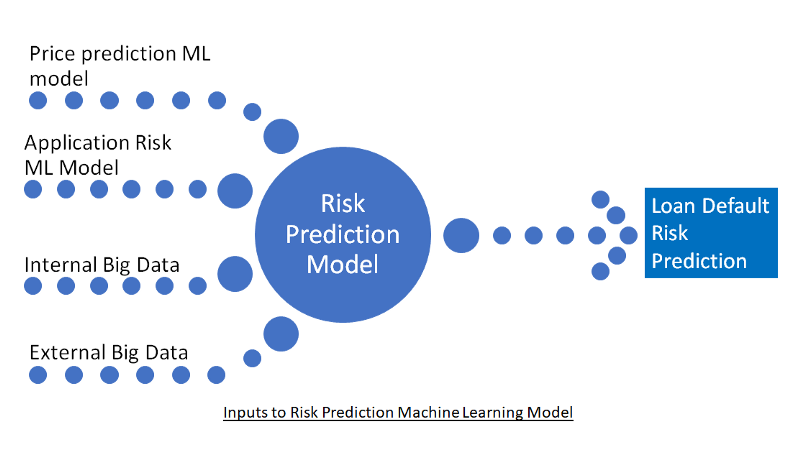
For a Loan Origination System, I recommend creating two categories of machine learning models:

**1. Initial Prediction Models:** These are the initial set of machine learning models we would need to create. The result of these models would act as an input to the final risk prediction model. Two examples of initial predictive models are:

1. Price prediction model: that predicts the price of the property for the duration of the mortgage. Historical prices of properties in an area are generally publicly available. We can look at these historical prices and create a model that predicts the price of the property in the future. If the prices in the area tend to go up, then the financial downside of loan default can be offset by the predicted increase in property prices. The output of the price prediction model would act as an input to the risk prediction model.

2. Loan application-based risk prediction model: a model that looks at the loan application data and documents, and makes a prediction on the risk of default. In order to make this prediction, we would create a model based on historical loan application data and documents of customers who had taken loans. Based on this historical data, we would create a model to predict the probability of loan default of the current loan application.

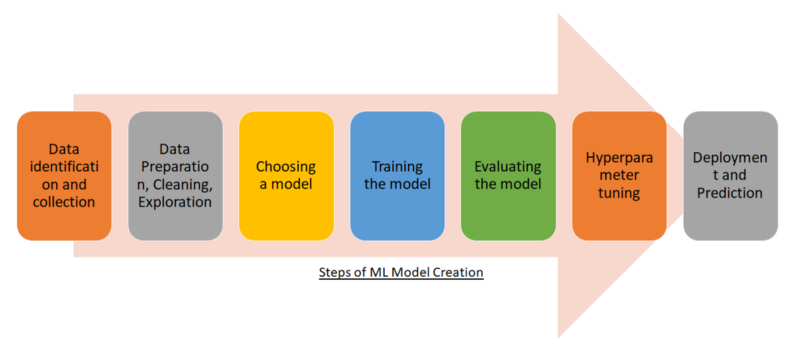
**2. Final Risk prediction model:** This model would look at all the features (input parameters) that are available in the bank’s BigData, including the output of the initial prediction models, and predict the risk of loan default. In the next section, I have specified what would constitute the bank’s BigData including internal Bigdata and external BigData. This model will help the lender assess the risk more accurately and completely than a human loan officer possibly could. This is because the loan officer would need to make the decision based on just the application forms and documents while the machine learning model would have additional access to BigData regarding the customer and the neighbourhood. The result of the Risk prediction model could be a scorecard prefilled that can be reviewed and used by the loan officer to make the decisions on the loan parameters. This process will also be a lot faster compared to a human being doing it.



The above three models will ideally be Supervised learning models. A supervised learning model is essentially a model that trains itself on historical data, with the intent to learn the correlation between input features (parameters) and the result, which in this case is loan default. Hence the data that we absolutely need to make this work is historical data on loans. We will split all the data we have into training data and testing data. Training data, which will be about 80% of all data, will be used to train our model so that it learns from history. We will then test our model on the remaining testing data, about 20%, to validate if our model is able to predict loan defaults accurately. If our results are not as good as we expect, we will tune the training parameters and input features, and redo these steps until we are happy with the results. Instead of a binary result of whether a customer will default or not, we will prefer a probability of default.

### Steps to create the ML model

Let us delve into the steps involved in creating such a machine learning model.



#### Step 1: Data identification and collection

We will need to identify all the data that exists in the lender’s system as a result of other activities that the customers perform with the lender e.g., banking, credit cards, previous loans, etc. Sufficient time and money will need to be spent on this step since the quality and quantity of the data will determine how well our prediction model will work.

We can call the different types of data collected as features, which is a term used in machine learning to identify input parameters for a machine learning model.

The data we would collect would fall into three categories:

**Internally available data:** such as neighborhood, age, income category, income sources, outgoing expenses, bounced cheques, credit card data, family data, family’s financial data, liabilities, past loan data, etc.

**Externally available data:** price trend for properties, future plans of development of the neighborhood, surrounding neighborhoods being developed that may impact the value of the property, plans for future public transport availability, government infrastructure being built around the neighborhood, new schools being built, heavy power lines being laid that may reduce the value of the property, factories planned that may cause pollution or increase demand in the area, etc.

**Borrower submitted data:** This is the data that was submitted with the mortgage application.

#### Step 2: Data Preparation, Cleaning, Exploration

In this step, we load all the identified data in a single BigData store. One of the key data element that we would need is past loan data for all customers, and the reason for that is explained in the above section. We then visualize the data to figure out correlations between different features, amongst themselves and with loan default data. Visualizations will also help us check for data imbalances, e.g. clusters on age group, neighborhoods, income category, etc. We want to make sure that in this step we remove imbalances by either removing the data that is causing imbalance or creating artificial data for the categories that are not well represented. This is also the stage where we split our data into training data and testing data.

We also clean the data. Some of the cleaning steps handle null values, outliers that can skew the results, deduplication of data, normalization of data, and correcting errors that may exist.

We may also want to reduce the number of features or create features to represent multiple features in this step. This is called dimensionality reduction. Principal component analysis (PCA) technique is often used for dimensionality reduction. Too many features make it harder to explain the prediction, hence investing time in this step is a good thing. In practice, we revisit this step as part of iterating the model after the model is created. By experimenting with dimensionality reduction, we often get better predictions.

#### Step 3: Choosing a model

At this step, we choose the type of model we will use. There are different types of supervised learning models. Some of the most used are:

1. Regression model

2. Classification model

3. Naive Bayesian model

4. Random Forest model

5. Support vector machine model

We will need to experiment with the different models to see which one gives the best prediction. I have personally seen the simplest of these, regression models, to be often most effective. Lasso and ridge regression should also be experimented with to see if they give good predictions since they offer benefits of dimensionality reduction.

#### Step 4: Training the model

In this step, we train the model on the training data and use it to predict using the testing data. This is a time-consuming step and involves a lot of incremental experiments with different parameters of the models and at times even different models. Since this step is often performed on large amounts of data, it is also very slow. The end result of this step is a model that does an acceptable job of predicting the data, which in our case is the probability of loan default.

#### Step 5: Evaluating the model

Here we use our trained model, and use it on the testing data. We look at how accurate the model prediction is in this step since this test is supposed to be a representation of the real world.

#### Step 6: Hyperparameter tuning

In this step, we try and improve the model by tuning the hyperparameters of the model. This is again a very experimental and time-consuming step. It is more of an art than science to tune a model.

#### Step 7: Deployment and Prediction

This is the final step where we deploy the model and start using it on real-time data. It is good to keep a retrospective watch on the performance of the model and retrain it if necessary since the real-world data can change and consequently the prediction accuracy of our model could degrade.

#### The end result of the ML model?

Once we have our model it will give us more informed risk about loan default. Based on this “more informed” risk, better interest rates can be offered to customers ensuring a better long-term relationship with the good customers. On the flip side, when the risk is higher, then the lender can protect itself by a higher interest rate, or reduced Loan to Value (LTV).

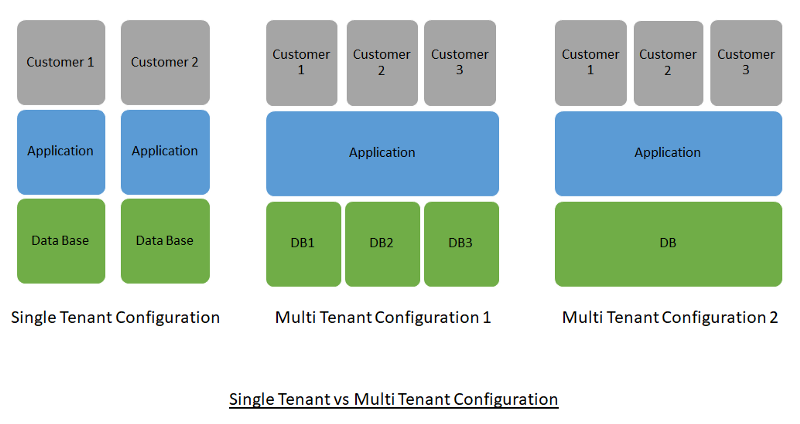
### What about the small lenders who don’t have BigData?

How can Loan Origination System providers give the benefit of BigData and Machine Learning to the smaller banks and building societies? Smaller banks have an issue where, unlike larger banks, they do not have access to years of customer data. The solution is offering a multi-tenant cloud SAAS offering.

#### Multi-Tenant Cloud SAAS offering

The LOS providers can offer a multi-tenant, SAAS-based cloud solution, where the customer insight can be leveraged by all the lenders, to the benefit of all the lenders. Let me explain that.

A single tenant solution would be where we have one bank, using a solution that is deployed on its dedicated cloud infrastructure. Now for a small institution, using such a solution on-premise or on the cloud with a dedicated infrastructure can be expensive. A shared cloud infrastructure would be a much more cost-effective solution.



Enter the multi-tenant cloud SAAS offering. Such an offering would be able to serve the software to multiple clients from the same shared cloud infrastructure. The data would also be separate for each client. This would reduce the cost for each customer.

Now, if the small lenders agreed to allow the machine learning model to look at all their combined data, and produce insight on the risk score of the borrower, there is a far higher chance of them getting more accurate predictions.

### Underwriters Using the ML model

Now that we have implemented a machine learning model that produces a scorecard based on the application data, the lender’s BigData, and the publicly available BigData, the underwriter can use their subjective judgment to validate the result. Over time, as the confidence in the results of the ML model grows, the entire process can be automated.

### Conclusion

In this article, we have seen how machine learning and Robotic process automation can be very useful tools that should be used by traditional Loan Origination Systems. The application processing step can be completed automated by RPA and hence save days or weeks of processing time. The underwriting process can be done a lot faster with a lot more confidence in the result with the help of machine learning. This will result in more competitive interest rates, more satisfied customers, and the bank is more protected from the risk of loan default.

### References

[1] <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-next-acronym-you-need-to-know-about-rpa>