Smith
SCHOOL OF BUSINESS
Queen's University

MMA 867: Predictive Modelling

Team Kipling - Assignment 3

Faye Ding	20250430
Matthew Gottlieb	20259474
Adaure Ibe	20254264
Angie Jung	20254266
Justin Liu	20254488
Mahendar Rawat	20254140

Due Date: June 4th, 2021

Table of Contents

Q2.

Business Problem

The main business problem of this project is regarding a new product (short-term line of credit) that a bank would like to offer to certain customers. Particularly, they would like to leverage historical data to develop a model to predict whether or not a customer will default on a loan. Naturally, the customers with the least likelihood to default would be the ones offered the line of credit.

The terms of the line of credit are a $25,000 offering for one month at 2% interest. This would ultimately bring in a profit of $500 for the bank (assuming the bank's cost of capital is $0), and they also believe that this would bring an additional lifetime value of $1,000. This would ultimately result in a total profit of $1,500 per recipient of the line of credit, assuming that they do not default. However, those who do end up defaulting bring a loss of -$5,000 as the bank would only recover $20,000 of the total $25,000 issued. Therefore, in order for this project to be feasible, the bank must ensure that the model can correctly maximize its True Negative ($1,500 value) predictions and minimize its False Negative (-$5,000) predictions.

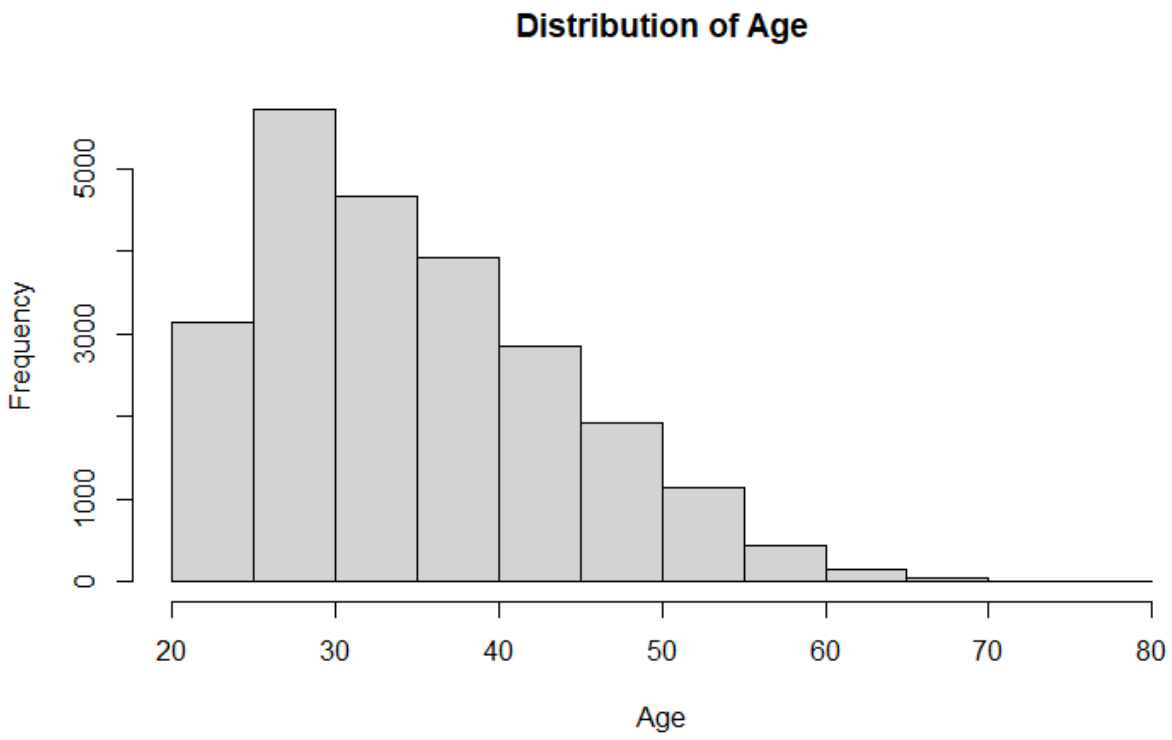
Exploratory Data Analysis

The first component of this project was to gain an understanding of the data through an exploratory data analysis. It is important to note that the data was received in two different files (historical data with default labels and prediction data without default labels). For the purposes of the EDA and the project as a whole, we used the historical data to gather insights and create models, while the prediction data was only seen at the end of the analysis to make final predictions.

The first question that came to mind during the EDA process was to determine what percentage of customers defaulted on their loans based on the historical data. Of 24,000 total records, we found 5306 occurrences of default, meaning that ~22.1% of total customers had defaulted on their loans. This percentage was startling to observe, as one would typically expect this number to be well under 5%. This also speaks to the importance of creating an accurate model, as the high probability and monetary loss associated with default in this circumstance could cause serious losses for the bank.

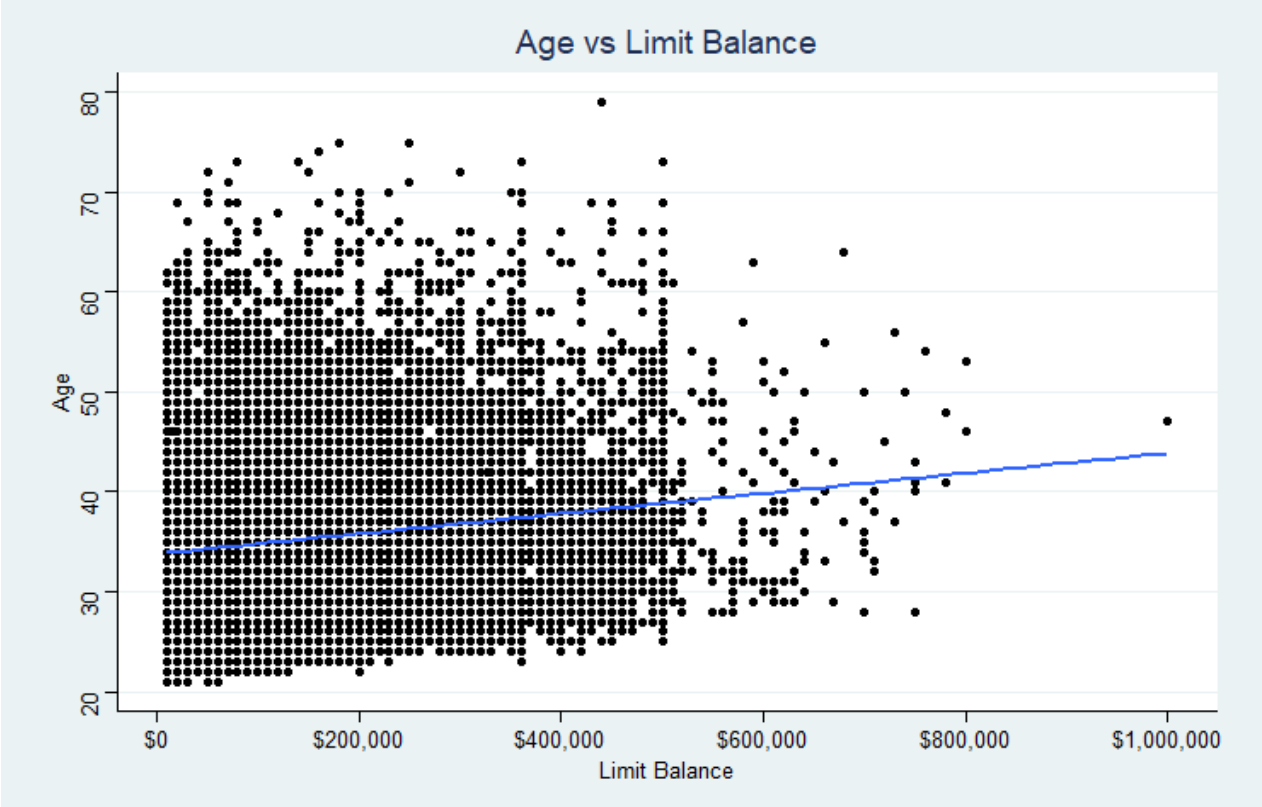
Another area of interest in the data surrounded the age variable. We wanted to gain an understanding of what the distribution of ages looked like among customers, as well as its summary statistics. This could then open the door to look at relationships that age shared with other variables, such as limit balance available, gender, and marital status. These are all insights that may help us to better understand who our customers are and ultimately pave the way for our predictive models to make decisions on which customers to grant loans to.

Below is the distribution of customers of customers by age:



It is visually apparent that the majority of customers fall within the 30-40 year old age buckets, with a mean age of 35.45 years old. We also observe a significant drop-off in customers past 50 years, which may be due to the fact that they are near retirement age and can rely on their own savings or investments to fund themselves. In terms of total dispersion, we see a minimum age of 21 years old and a maximum of 79.

The next question we wanted to gain insight into was how strongly age correlates with limit balance. Specifically, we wanted to know if the limit balance increased with age or if there was no relationship at all. Our initial assumption was that typically, as one gets older, their salary and net worth rises, leading to higher credit limits from the bank. When digging into the data, we found this relationship to be positively correlated but of relatively weak strength, with an r-value of 0.142. The below graph illustrates this relationship:



This graph highlights the relatively weak positive relationship that was touched on above. It is also visually apparent that the limit balance for the majority of customers is below $500,000. However, there are some outliers with limits as high a $1,000,000.

Finally, we will keep an eye on the importance of the Repayment Status, Bill Amount and Pay Amount variables. These variables are seemingly significant, as they indicate whether or not the customer is behind on their repayment status and what their actual bill amounts are. We anticipate these variables to play a crucial role in the predictive power of our machine learning model, as they essentially speak to the reliability one has in repaying their debts.

Data Preparation

When reviewing the structure of the data itself, we identified a few areas in need of cleaning/preparation. Our first step was to resolve issues within the Education and Marriage categories. In Education, the data dictionary noted that levels 5 and 6 corresponded to “unknown”, while level 4 represented “other”. As we could not find any distinction between the three categories, we merged the values of 5 and 6 into 4 to create one category the represented unknowns. We also observed some values that were represented by category 0. As this was not part of the data dictionary, we treated this as missing data and imputed these values using the mode later on. We noticed a similar issue of category 0 within the Marriage variable and followed the same course of action to treat these observations.

Additionally, we also had to change the actual object structure of some of our variables. For example, we converted our target variable, default_0, to a factor to represent the two levels of the outcome. Similarly, this was also done for the six levels of the Pay variable. It is important to note that we added 3 to each value in the Pay variable in order to ensure each value was positive.
Another transformation was made to the education variable. This was transformed into an ordinal factor to represent the ordered difference in education classes. More specifically, graduates had the highest ranking, followed by undergraduate, high school and others.

Furthermore, we decided to centre and scale our data. This is primarily due to the fact that our numeric data has significant differences in scale. For example, we have ages between 21-79, but limit balances in the thousands and even millions. Doing this would allow our data to be centred around a mean of 0 and ensure that these variables are around the same scale of order, improving machine learning interpretability for our modelling efforts later on.

We also decided to apply a Box-Cox transformation to our data. As we knew that we would be testing methods that assume normality, transforming our data accordingly would ensure that the algorithms function as intended.

Feature Engineering

Below is a summary of the features we added into our data:

- **EDUCATION_GRP**: creating 4 groups for education levels; 1 as Graduate, 2 as Undergraduate, 3 as Highschool, 4 as others.
- **MARRIAGE_GRP**: creating 3 groups for marital status; 1 as Married, 2 as Single, and 3 as Divorce.
- **AGE_GROUP**: creating 6 age groups for the AGE_GROUP variable; Under 25, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 and above.
- **BILL_AMT**: creating variables of BILL_AMT_MEAN, BILL_AMT_MEDIAN, BILL_AMT_SUM, BILL_AMT_MIN, BILL_AMT_MAX, and BILL_AMT_SD. For these 6 variables of bill payment amount, we have created 6 new variables in which we calculated the mean, median, max value, min value, standard deviation, and sum value, respectively.
- **PAY_AMT**: creating variables of PAY_AMT_MEAN, PAY_AMT_MEDIAN, PAY_AMT_SUM, PAY_AMT_MIN, PAY_AMT_MAX, and PAY_AMT_SD. For these 6 variables of bill payment amount, we have created 6 new variables in which we calculated the mean, median, max value, min value, standard deviation, and sum value, respectively.
- **LIMIT_BAL_CAT**: creating 5 balance limit groups for this variable; less than 50k, 50k to 100k, 100k to 250k, 250k to 500k, 500k plus.
- **PAY_CAT**: for the PAY variable we created PAY_CAT with 6 groups; for PAY variables of -2, we group them as No Credit, -1 as Pay Full, 0 as Revolving Credit, 1, 2, and 3 as

Delay1_2_3 (delay for 1, 2, 3 months), larger than and equal to 4 as Delay4plus, and others.

- **DIST_LIM**: another new variable created is DIST_LIM which is the percentage of difference between Bill Amount and Limit Balance.
- For variables of EDUCATION_GRP, MARRIAGE_GRP, AGE_GRP, LIMIT_BAL_CAT, PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, and PAY_CAT, we set them as dummy variables.

Final dataset: After creating the features above, we column bind the original dataset and the dataset with new features and take out the existing variables of AGE, MARRIAGE, EDUCATION, and SEX. After these sub-setting steps, we finalize our ultimate dataset for the modelling procedure.
In addition, we randomly split the data into 3 sets (training, validation, and testing) using default_0 for stratification to build our predicting model. For the training dataset, we used 22000 rows of data; and for validation and testing datasets, we assigned 1000 rows of data for both datasets. On the validation dataset, we will run cross-validation on these 1000 rows of data; on the test dataset, we will test our model.

Model Training & Selection (with *tidymodels*)

a) Model Building

The process of building model in *tidymodels* is as follows:

1) **Recipe**: We started with the *recipe* function from the *tidymodels* package to declare our target and the predictor variable(s) from the training dataset. As we have learned previously during the data exploration stage, the ID in the new applicant's data is in a different format and can't be used as a predictor variable.

2) **Cross-Validation:** As a next step, we set up the parameters for cross-validation, which we would use to tune the hyper-parameters and test our model's ability to predict new data. We have already created a validation dataset with 1000 observations, which we use for k-fold cross-validation to build a set of 10 validation folds with the function *vfold_cv.*

3) **Model Specification:** We decided to start by building some of the most used models for solving a classification problem. We would then choose a model with the best performance on the test dataset. We chose to pick Logistic regression, Random forest, K-nearest neighbor and Decision Tree model types and accordingly specified these models in R.

4) **Workflow:** We used the package *workflow* to bundle the recipe and the models we specified earlier.

5) **Hyper-parameter tuning:** Before we use the models we have specified, we have to tune the hyper-parameters so that the model can optimally solve the problem. We used a grid search approach and provided a range of values for different hyper-parameters and then picked the one which returns the highest accuracy.

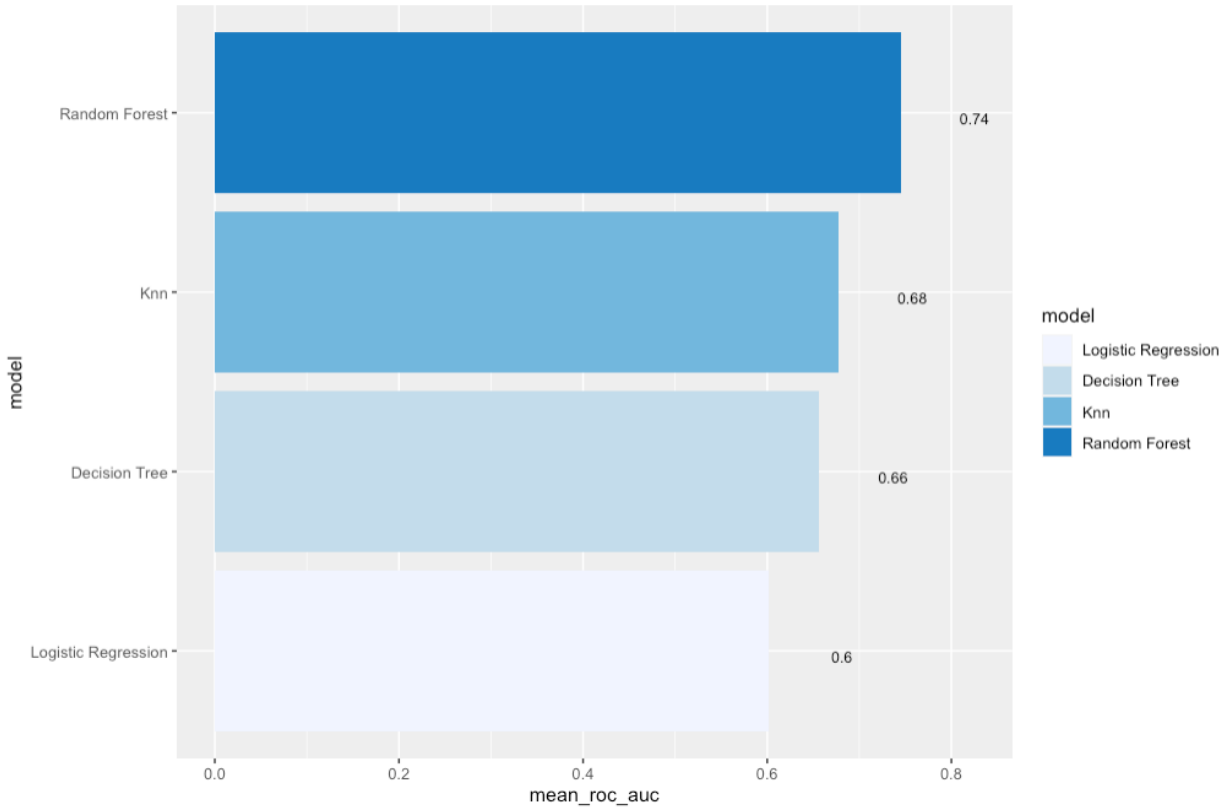
6) **Model training:** At this point, we bundled recipe and models with workflow to train the models, as well as we use validation data to estimate the performance of the model using the *fit_samples()* function and store the results. We collect the performance metrics to compare the models.

b) Evaluate Models

Once all the models are trained, we extract the metrics and compare them to find the best model with highest Area Under the Curve (AUC).

From the below table we can see Random Forest is the best model.

	model	mean_accuracy	mean_f_meas	mean_kap	mean_precision	mean_recall	mean_roc_auc
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Logist…	0.715	0.812	0.206	0.824	0.805	0.601
2	Random…	0.796	0.879	0.253	0.817	0.951	0.745
3	Knn	0.735	0.830	0.224	0.829	0.832	0.678
4	Decisi…	0.786	0.872	0.233	0.816	0.937	0.657

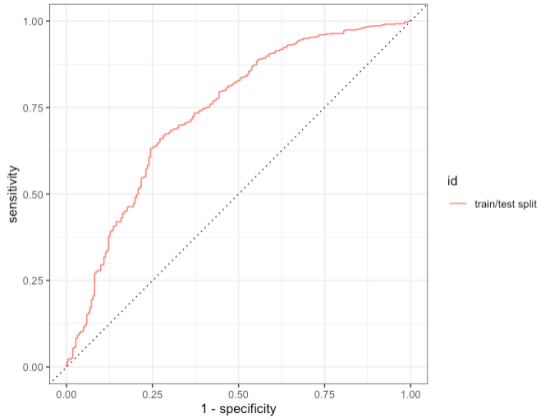


Evaluate on Test Dataset

After choosing the best performing model on the validation dataset, using the *last_fit()* function we applied the model to the testing dataset to assess the performance of the model.

Our model has almost similar performance on the test data.

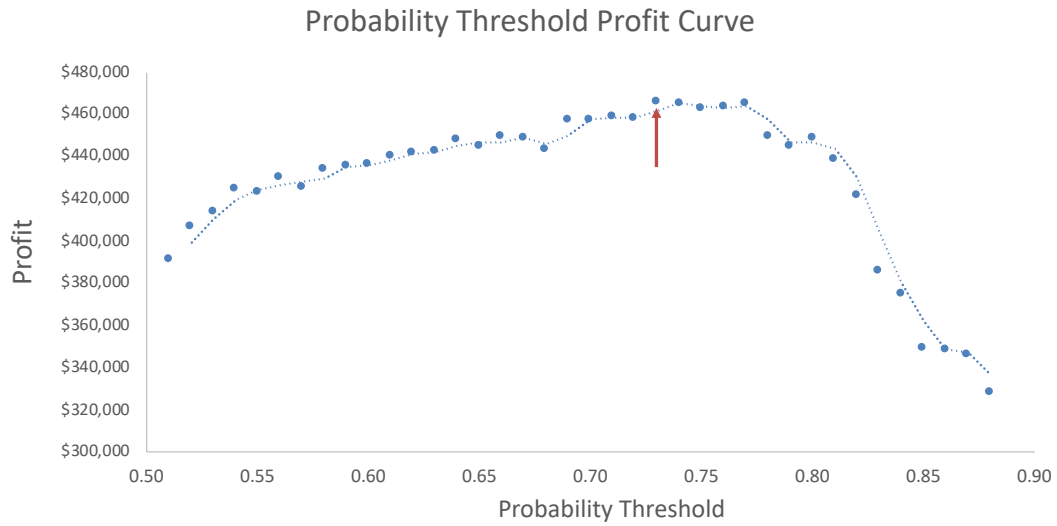
.metric	.estimator	.estimate	.config
<chr>	<chr>	<dbl>	<chr>
1 recall	binary	0.960	Preprocessor1_Model1
2 precision	binary	0.821	Preprocessor1_Model1
3 f_meas	binary	0.885	Preprocessor1_Model1
4 accuracy	binary	0.806	Preprocessor1_Model1
5 kap	binary	0.283	Preprocessor1_Model1
6 sens	binary	0.960	Preprocessor1_Model1
7 spec	binary	0.262	Preprocessor1_Model1
8 roc_auc	binary	0.737	Preprocessor1_Model1



One thing to note here is that the specificity is quite low with the model, primarily because we have not tuned the probability threshold yet. We know that the probability of someone defaulting on credit is not 50%; it should be far lower than that. During data exploration, we learned that only approx. 22% of customers defaulted on their bill payment.

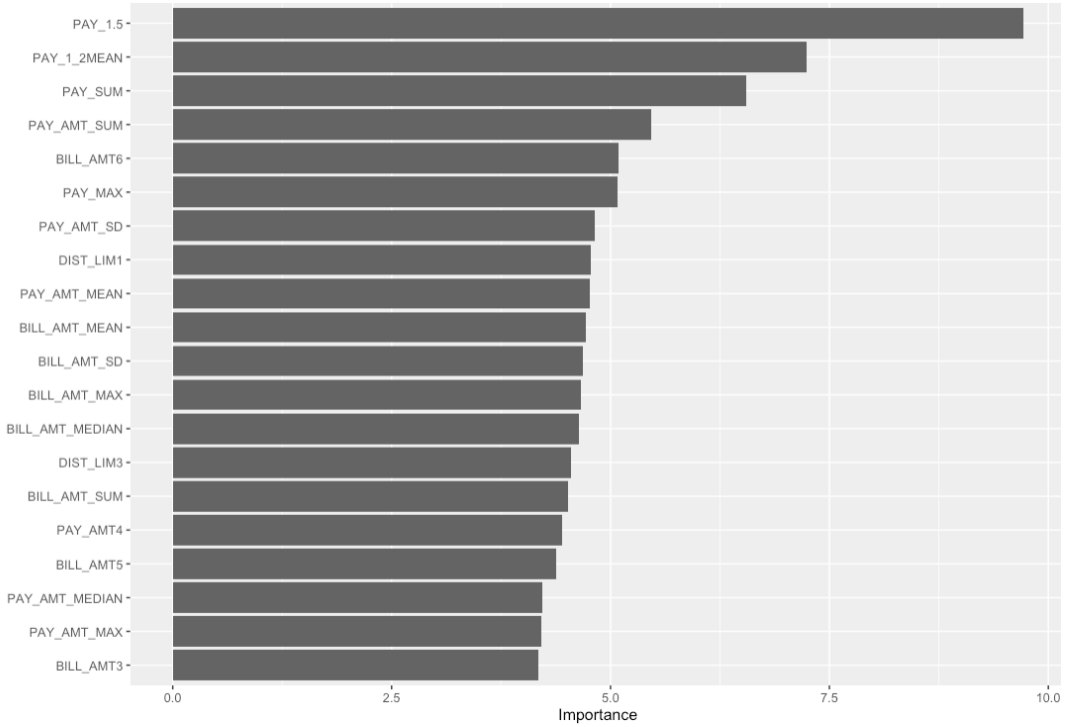
However, rather than tuning the probability threshold to maximize the accuracy of the model, our goal is to maximize the profit. We know that for every True Positive (credit issued and repaid), the bank earns a profit of $1500. On the other hand, for every False Positive (credit issued and not repaid), the bank bears a loss of $5000.

The below chart depicts the profit value at a range of probability threshold for not defaulting on credit. We can see that at 0.73 cut-off, we would make the maximum profit.



After using the new probability cut-off, we observed that the specificity has gone up even though the accuracy has decreased slightly. We should be ok with that because our goal is to maximize the profit not the accuracy at this point.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 637  98
         1 142 123

               Accuracy : 0.76
                 95% CI : (0.7323, 0.7862)
    No Information Rate : 0.779
    P-Value [Acc > NIR] : 0.930306

                  Kappa : 0.3494

 Mcnemar's Test P-Value : 0.005509

            Sensitivity : 0.8177
            Specificity : 0.5566
         Pos Pred Value : 0.8667
         Neg Pred Value : 0.4642
             Prevalence : 0.7790
         Detection Rate : 0.6370
   Detection Prevalence : 0.7350
      Balanced Accuracy : 0.6871
```



Last but not the least, we also wanted to look at the variables which are very important in this model. We observed that most of the new variables we feature engineered are at the top, but here is the list of the important variables that these new features are derived from:

1. Pay Variables (PAY_1, PAY_2, PAY_3, ..)
2. Bill Amount (BILL_AMT_1, BILL_AMT_2, ..)
3. Pay Amount (PAY_AMT_1, PAY_AMT_2, ..)
4. Limit Balance (LIMIT_BAL)

Prediction on New Applicants

At this stage, we are ready to apply our final model to the new applicant data to predict which customers should be issued a credit.

Using the ***predict*** function, we extracted the probability values and then, using the probability threshold, we classified 1s to the customers who should be issued a credit and 0s to customers who should not be given credit.

Q3.

(a):

Since we already have our model's performance with "SEX" variable, we re-trained the same model without the gender column. We observed that the accuracy of the model dropped by approximately 4%.

Confusion Matrix and Statistics:

Model with Gender Variable

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 600  87
         1 179 134

               Accuracy : 0.734
                 95% CI : (0.7055, 0.7612)
    No Information Rate : 0.779
    P-Value [Acc > NIR] : 0.9997

                  Kappa : 0.3277

 Mcnemar's Test P-Value : 2.411e-08

            Sensitivity : 0.7702
            Specificity : 0.6063
         Pos Pred Value : 0.8734
         Neg Pred Value : 0.4281
             Prevalence : 0.7790
         Detection Rate : 0.6000
   Detection Prevalence : 0.6870
      Balanced Accuracy : 0.6883

       'Positive' Class : 0
```

Model without Gender Variable

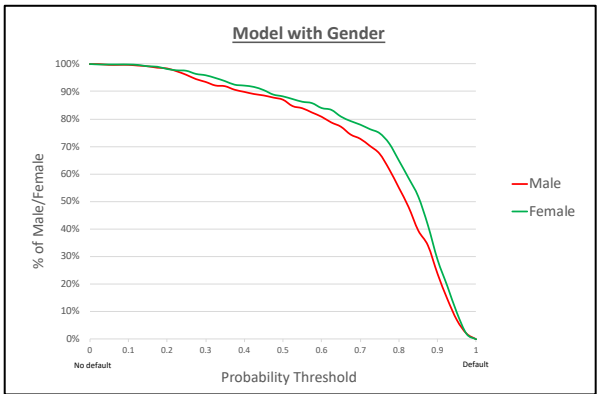
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 550  78
         1 229 143

               Accuracy : 0.693
                 95% CI : (0.6634, 0.7215)
    No Information Rate : 0.779
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2837

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7060
            Specificity : 0.6471
         Pos Pred Value : 0.8758
         Neg Pred Value : 0.3844
             Prevalence : 0.7790
         Detection Rate : 0.5500
   Detection Prevalence : 0.6280
      Balanced Accuracy : 0.6765

       'Positive' Class : 0
```

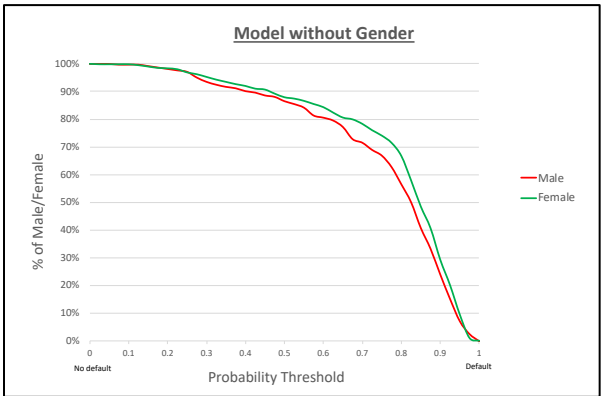
(b):

We observed that the percentage of males vs. females defaulting thus not getting credit is very similar in both models. The model predicts that females have a higher percentage of not defaulting on their credit than males. Still, that ratio stayed consistent more or less in both models.

Let's look at the results of both models. The model without gender has an increase of 6% in True Positives, but that comes at the cost of 26% increase in False Positives. Here 0 indicates not default, and 1 is for default, which means the bank would lose on the opportunity cost from issuing credit to customers who would have repaid. On the other hand, we also see a decrease of False Negatives from the model without gender, but at the cost of a substantial decrease in True Positives. As we have learned from the

project details that bank makes $1500 for every True Negative (predicted- not default/credit issued, truth – not default/repaid) and lose $5000 for every False Negative (predicted – not default/credit issued, truth – default/not repaid). Based on these details, we estimated that the bank would lose around $30,000 using the model without the gender variable.





(c):

When exploring the implications that algorithmic decision-making can have in terms of equality, anti-discrimination and ethics, there are certainly multiple perspectives that can be had. This topic has gained a lot of traction in recent years, as we increasingly rely on AI to make decisions. We believe that the key questions to be asked are 1) how does one guide an AI not to make unfair discriminatory decisions, and 2) How does leaving out variables that typically have significant predictive power (Gender, Age) impact the bottom line of businesses that use these tools?

When thinking about the basis of predictive modelling (or even human decision-making!), almost all decisions are made with discrimination. For example, a linear regression model to predict the price of a car will "discriminate" between variables and place higher importance on those that it will minimize the error and maximize the reward. Similarly, a university admissions board will "discriminate" against applicants on the basis of granting admission to only those students they believe will succeed in the program. However, there is a large ethical debate regarding *which* variables the decision-making process is able to discriminate against. In our particular case of granting lines of credit, we focused on the difference in predictive power between a model that included gender and one that did not. Specifically, we noticed that the model trained on data that contained the gender variable had stronger predictive power than the model trained without it. As mentioned in the case itself, businesses operating in some countries do not have the option of using the gender variable or even collecting it in their data. An example of this would be the US with their Equal Credit Opportunity Act (ECOA) that prohibits the collection of characteristics including gender, race, marital status, etc. Conversely, companies in the EU are allowed to collect this information; however, as of 2011 are barred from including it in training their models. And finally, in Canada, it is noted that there is a lot of "wiggle room" in which these variables could potentially be included.

Due to the fact that it is not illegal to collect this information in the EU and Canada, it is entirely possible that this information can make its way into predictive models indirectly. For example, it is possible that

one's gender can somehow be tied to a separate variable (i.e. digits in a client's ID can represent characteristics of the customer), and therefore influence the performance of the model. This would enable businesses who use a grey-area approach to potentially bypass regulatory checks, as these sensitive variables would not explicitly be mentioned in the data dictionary/metadata. Conversely, companies in the US that deploy predictive models do not even have this option. By not collecting this information in the first place, businesses have no way of adding these variables to their predictive models. As many see it, putting laws in place to explicitly prohibit the collection of information about gender, race etc. is the only way to truly stop predictive models from building in unfair discrimination to maximize results, as they would not have that option to begin with.

The other side of the debate relates to the business implications of using or not using these variables. As noted earlier, there was a clear difference in performance when comparing the models trained with and without gender. Based on our results, we noticed an average profit of $465,000 when deploying the model trained with gender, as well as $435,000 without gender. This is a difference of -$30,000 and approximately a 6.5% decrease. As this was only based on a pool of 1,000 customers, one can quickly infer the difference in profits generated between models if a project like this were to be executed on a larger scale. It also becomes obvious why a company operating in a country with more lenient laws would be keen on collecting and including as much information as possible about their customers for use in their predictive models. Additionally, there is a further debate to be had about the ethics of the outcomes of models that purposely leave out information. For example, if a particular customer received a favourable interest rate based on their individual characteristics (not just gender, but reliability in repayment etc.) and suddenly had it increased due to the removal of some variables, they would not be pleased. Conversely, some customers may receive the opposite outcome. The key point is that removing characteristics that allow models to uniquely identify customers has the potential to both negatively/positively impact them, but it ultimately will view them as more of a generalization and less of an individual. From a business perspective, this has to be carefully managed as it has the capacity to damage customer relationships depending on the outcome.

Overall, the debate of the use of characteristics, including gender, age, marital status, etc., in predictive models is contentious. On the one hand, many activists believe that including these variables in predictive models can result in discriminatory outcomes deemed to be unfair. On the other hand, including as much information about the customer as possible will most likely increase the model's predictive power and ultimately create higher rewards for its owner. Whether or not a business decides to employ these methods is up to the discretion of the country's regulatory laws it operates in and its moral high ground. Ultimately this is an issue that will certainly continue to be a hot topic for debate and will be intriguing to monitor for years to come.