MAT012 Credit Risk Scoring

ASSIGNMENT 2022-2023

KARAN MANOHARAN C22070780

Part A:

Topic 1: Critically examine what needs to be considered when developing a credit risk scoring model.

A credit risk model is a statistical model used by lenders and financial institutions to assess the likelihood of a borrower defaulting on a loan or credit obligation. Credit risk models use historical data on borrower behavior and other variables to estimate the probability of default or other credit-related events. These models are typically used to determine the creditworthiness of borrowers and to make decisions about whether to grant credit, how much credit to grant, and at what interest rate. Credit risk models may use a variety of statistical techniques, such as logistic regression, decision trees, or neural networks, to analyze data and estimate credit risk. The development of credit risk models is a complex process that requires careful consideration of various factors, such as data quality, model assumptions, and regulatory requirements.
The three aspects which influence a borrower's credit risk:

1.The probability of default (abbreviated POD or PD): It indicates the likelihood that the consumer will be unable to make expected debt payments. The default risk of each borrower is best represented by a combination of two factors: credit score and debt-to-income ratio. Higher PODs are associated with higher interest rates and required loan deposits on average. Consumers can help share the risk of loan default by providing collateral. (Ross 2021).

2.Loss given Default: The total loss incurred by the creditor if the loan is not settled. This is an important aspect of credit risk modelling. The credit risk profiles of two borrowers with the same credit score and debt-to-income ratio will be very different if one borrows significantly more than the other two. This is because the lender's loss is significantly greater when the amount is larger in the event of a default. Again, interest rates and down payments are significantly affected. The interest rate will be significantly affected if the borrower is willing to put up collateral (Anukrati Mehta 2019).

3.Exposure at Default (EAD): Exposure at default (EAD), like loss given default, is a calculation of a creditor's overall potential losses at any given point in time. Even though EAD is almost always associated with a bank, total exposure is a critical concept for anyone with extended credit. The concept of EAD is that outstanding amounts that can accumulate prior to a default determine potential risk. For debts with credit constraints, such as credit cards or lines of credit, risk exposure assessments should consider not only current amounts, but also the possibility of account balance increases before the borrower defaults (Ross 2021).

Credit Risk Scoring typically involves building models – scorecards – that estimate the risk of default from a sample of past customers and their observed repayment behavior.
The score that's created for a customer by a lender is compared to a cut-off value:

- if the score >= cut-off value, credit is granted.
- if score < cut-off value, credit is denied

The cut-off value must be decided by the lender.
There are two types of decision in granting credit, with much in common from a modeling perspective:

Whether credit should be granted to the new applicant- This is called application scoring.
Whether credit should be extended to an existing account. -This is called behavioral scoring.

Credit Scoring – Application Scoring

•Estimates default risk at the time the applicant applies for the loan – the score is a representation of the applicant's default risk
•Uses a predetermined definition of what is meant by default
•Uses application variables – age, sex, years at address, income, capital, collateral, repaying ability, past data, and present market conditions.
•It also uses Bureaus-credit reference agencies (CRAs), e.g.: Experian – experian.co.uk, Equifax – equifax.co.uk, and Transunion – transunion.co.uk
•These agencies calculate credit scores for a consumer based on the borrowing and payment history of that consumer.

Credit Scoring – Behavioral Scoring

•In behavioral scoring, a company will be looking at assessing risk for existing customers through internal behavioral data
•It will involve looking at risk over some time window (e.g., 3 months, 6 months, 12 months) of an existing customer given their recent behavior
•That behavior may include:
-Any missed payments.
-Trends in certain internally-held variables – for example, change in current account or savings account balances
-Changes to scores from the CRAs – are there emerging risks for an existing customer because of a deterioration in accounts they hold elsewhere?

Most Credit assessments seemed to be built on the belief that the 5Cs were all that mattered - character, capital, collateral, capacity, and condition.

The sections that follow go into detail on the various steps that comprise the process of developing the credit risk scoring model. Before beginning actual modelling work, it is recommended to examine the sample data. Simple statistics such as the distribution of values for each characteristic, mean/median values, proportions of missing values, and value distributions can provide valuable business information, and examining them is an excellent exercise for ensuring data accuracy. Furthermore, data should be examined for interpretation (for example, to ensure that "0" refers to zero rather than missing values) and to ensure that any unusual values, such as 99 or 999, are properly recorded. This stage confirms that the data was gathered as directed and that all its components, including data, were collected.

Most financial industry data contains missing or inconsistent values for a given attribute. These could be mis-keyed values, outliers representing extreme cases, or fields that were not collected, terminated, unavailable, or not completed by applicants. Logistic regression requires complete datasets with no missing values, whereas other statistical techniques, such as decision trees, are unconcerned about missing values. To handle missing values, one of three options exists: either

remove all data with missing values from the model, exclude characteristics or records with a significant number of missing values, or statistically impute missing values. Outliers are values that are significantly outside of a given attribute's normal range. These figures are typically removed because they could have a negative impact on the regression results. There is correlation, and it needs to be dealt with. Prior to the regression stage, but either before or after the initial characteristics analysis, is the correlation step.

The first characteristic analysis involves two major tasks. The first stage is to assess each trait's ability to predict performance on an individual basis. This method, also known as univariate screening, is used to eliminate untrustworthy or irrational attributes. The strongest traits are then used to form groups. This is true for the properties of both continuous and discrete characteristics. Binning is the process of organizing continuous (ungrouped) characteristics into bins to create scorecards based on those characteristics. Once the strongest traits have been aggregated and ranked, variable selection is complete. To determine how strong a trait is, four basic factors are used:

• The ability of each attribute to predict. For this, the weight of evidence (WOE) metric is employed.

• The distribution and trend of evidence weight across a characteristic's grouped qualities.

• The characteristic's tendency for prediction. This is done using the Information Value (IV) measurement.

• Business and operational considerations

Some analysts perform additional variable selection algorithms (such as those that rate predictive power using Chi Square or R-Square) prior to classifying attributes (Siddiqi 2006).

Using a variety of predictive modelling techniques, a group of attributes with the highest predictive potential can be chosen when creating the preliminary scorecard. Among the methods used in the industry are neural networks, decision trees, and logistic regression. Regardless of the modelling approach used, this procedure should yield a scorecard with the optimal combination of traits, considering factors such as:

- Correlation between characteristics
- Final statistical power of the scorecard
- Interpretability of characteristics at the branch/adjudication department
- Implement ability
- Transparency of regulatory requirements methodology

The analysis of a previously rejected application to extrapolate their behavior is known as reject inference. There will be some goods that have been rejected, just as there will be some bads in the population. This procedure recreates population performance for a 100% approval rate, making the scorecard preparation process more relevant (Siddiqi 2006).

The post-inferred dataset is then subjected to the same initial characteristic analysis and statistical methods to generate the final set of characteristics for the scorecard (such as regression). You are not limited to the qualities identified in the preliminary scoring at this stage. The process of selecting characteristics must consider the entire development dataset because some qualities may appear weaker and others stronger after reject inference (Siddiqi 2006).

Most scorecard developers would create two or three different scorecards to complete any project. Given the level of control and flexibility involved in the development technique, creating multiple score cards become a more appealing option. We must answer two questions to choose one of them as the winner: Which scorecard is the most accurate? How reliable is the scorecard? The answers to the questions are provided by statistical and business metrics. Scorecards are used to forecast whether a case will be successful or unsuccessful. They are also used as prediction models to distinguish between good and bad cases. Statistics on misclassification are a useful tool for assessing whether a scorecard is offering the appropriate differentiation. To assess the severity of such misclassification and to compare various scorecards, a variety of metrics are used. These measurements contrast the actual and expected quantities of good and bad for a given cut-off. As it can be seen below, the measures are based on a confusion matrix.

It would be preferable to have a scorecard with the highest percentage of "true" cases and the lowest percentage of "false" cases (Siddiqi 2006). To assess the misclassification, four key metrics are used:

		Predicted	
		Good	**Bad**
Actual	**Good**	True Positive	False Negative
	Bad	False Positive	True Negative

Fig 1.1: Confusion Matrix

- Accuracy: (true positive and negative)/ (total cases)
- Error rate: (false positive and negative)/ (total cases)
- Sensitivity: (true positives)/ (total actual positives)
- Specificity: (true negatives)/ (total actual negatives)

After the final scoring has been determined, the modelling findings must be verified. Validation ensures that the model has not been overfitted and that it is appropriate for the target population. The scorecard is considered valid if there is no discernible difference between the two sets of data. To accomplish this goal, two curves are typically visually examined. Any measure of goodness of fit, such as information value or the Least Squares approach, can, however, be used. A second method of validation is to compare the development statistics or divergence statistics for the development and validation samples. Validation can also be performed by comparing the good/bad ratio by score range for the development and validation samples. Further research needs to be done in the above-mentioned methods.

Topic 2: Explain how, in theory, Cox's proportional hazard model for survival analysis can be used for constructing a scorecard. Comment on the relative popularity of Cox's PH model versus logistic regression in scorecard construction.

Survival analysis is a statistical technique for calculating the time until an event of interest occurs. It is commonly used in medical research, engineering, social sciences, and business to study the time it takes for an individual to experience a particular specific event or outcome, such as the death of a person, disease, mechanical part failure, or customer churn. In survival analysis, the result of interest is often referred to as a "failure event" or "event of interest," and the time until the event occurs is referred to as "survival time."
Survival analysis accounts for the fact that not all participants in a study will witness the event of interest, and some may be censored (i.e., their survival time is not fully observed). Censoring is a form of missing data problem in which time to event is not observed for reasons such as termination of study before all recruited subjects have shown the event of interest or the subject has left the study prior to experiencing an event. Censoring is common in survival analysis.
Given the characteristics of the individuals in the study, the goal of survival analysis is to estimate the probability of experiencing the event of interest over time.
Survival analysis is commonly performed using specialized statistical models such as the Kaplan-Meier estimator, Cox proportional hazards model, and parametric survival models. These models can be used to calculate survival probabilities, hazard rates, and other relevant parameters, as well as to identify the factors that influence the likelihood of encountering the event of interest.
Survival analysis has numerous applications, including healthcare, epidemiology, finance, and marketing. (Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer, 89(2), 232-238. https://doi.org/10.1038/sj.bjc.6601118)

The Cox Proportional Hazards (PH) Model is a commonly used method in survival analysis. It is a semi-parametric model that allows for the estimation of the hazard function, which is the probability of an event occurring at a specific time given that the event has not yet occurred. The Cox PH model assumes that the hazard function is proportional across different levels of the covariates, meaning that the hazard ratio between two groups is constant over time.
The Cox PH model is useful in survival analysis because it allows for the modeling of the time to an event, such as death or default, rather than just whether an event occurs. This is particularly relevant in credit risk analysis, where the time to default is of interest.
In the Cox PH model, the hazard function is modeled as a product of a baseline hazard function and a function of the covariates. The baseline hazard function is non-parametric and represents the hazard in the absence of any covariates. The function of the covariates is modeled as a log-linear function, which assumes that the effect of the covariates on the hazard is constant over time.
However, it is often the case that the proportional hazards assumption does not hold, and the effect of the covariates on the hazard changes over time. In such cases, splines can be used to model the covariates in a more flexible way. Splines are flexible functions defined by piecewise polynomials that are joined in points called "knots".
In summary, the Cox PH model is a useful method in survival analysis that allows for the modeling of the time to an event of interest. It assumes that the hazard function is proportional

across different levels of the covariates, but this assumption can be relaxed by using splines to model the covariates in a more flexible way.(Lore Dirick, Gerda Claeskens and Bart Baesens.ORSTAT, Faculty of Economics and Business, KU Leuven, Leuven, Belgium; Leuven Statistics Research Center (LSTAT), KU Leuven, Leuven, Belgium; LIRIS, Faculty of Economics and Business, KU Leuven, Leuven, Belgium; and School of Management, University of Southampton, Southampton, UK)

To build a scorecard using a Cox model, you would typically follow these steps:

- Data preparation: Collect and clean the data and prepare it for analysis. This may involve selecting relevant variables, dealing with missing data, and transforming variables as needed.

- Model development: Fit a Cox proportional hazards model to the data, using the selected variables. This involves estimating the coefficients for each variable, which represent the effect of that variable on the hazard of default.

- Scorecard development: Once the model is developed, you can use the coefficients to create a scorecard. This involves assigning points to each variable based on its coefficient, and then summing the points to get a total score for each borrower.

- Validation: Validate the scorecard to ensure that it performs well on new data. This may involve testing the scorecard on a holdout sample or using cross-validation techniques.

- Implementation: Once the scorecard is validated, it can be implemented in a credit scoring system to help make lending decisions.

A scorecard can be built using the Cox proportional hazards model, which is a widely used statistical method in survival analysis. The Cox model allows for the estimation of the hazard ratio, which is the ratio of the hazard rates for two groups with different levels of a predictor variable. It is important to note that the Cox model assumes that the hazard ratio is constant over time, which may not be the case in all situations. In addition, the scorecard should be developed and validated using appropriate statistical techniques and should be interpreted with caution, considering the limitations of the underlying data and model. The Cox model scorecard is a powerful tool for risk prediction, and it has been used in a variety of clinical settings. For example, the scorecard has been used to predict the risk of death in patients with cancer, the risk of disease progression in patients with heart disease, and the risk of treatment failure in patients with HIV/AIDS.

Both the Cox's proportional hazards (PH) model and logistic regression are commonly used in scorecard development, but they serve different functions.

Logistic regression is a widely used method for modelling binary outcomes (for example, default vs. non-default) and estimating the likelihood of an event occurring given a set of predictors. It is commonly used in credit scoring to predict the likelihood of a borrower defaulting on a loan.

Because it can handle both continuous and categorical predictor variables, logistic regression is especially useful in credit scoring. Furthermore, logistic regression can handle predictor variable interactions, allowing for more complex modelling. Interactions between income and debt-to-income ratio, for example, can be included in the model to capture the effect of income on default risk at various debt-to-income ratio levels.

Cox's PH model, on the other hand, is used for modelling time-to-event data (for example, time to default), where the hazard (i.e., the probability of the event occurring at any given point in time) is modelled as a function of predictor variables. In survival analysis, this model is commonly used to estimate the effect of predictor variables.

Logistic regression and linear regression are two commonly used statistical methods in credit scorecard development. While both methods aim to create a predictive model that assigns a risk score to each customer, their assumptions and applicability to different types of data differ.

Linear regression is a statistical method for modelling the relationship between one or more predictor variables and a continuous outcome variable. Linear regression can be used in credit scoring to create a model that predicts the likelihood of default based on the borrower's credit history and other relevant information. Linear regression, on the other hand, is not always appropriate for credit scoring because it assumes a linear relationship between the predictor variables and the outcome variable. In many cases, where the relationship is more complex or non-linear, this assumption may not hold true.
While all models can be used to build scorecards, their popularity is determined by the type of data being modelled. When the outcome of interest is binary, logistic regression is more commonly used, whereas Cox's PH model is better suited for modelling time-to-event data.

In conclusion, the popularity of Cox's PH versus logistic regression in scorecard construction is determined by the type of data being analyzed and the type of outcome being predicted.

Topic 3: A lender would like to extend its ability to offer credit to those with lower credit scores and is considering doing this through a combination of risk-based pricing and the use of more and different data in its credit scoring model. Discuss the implications of these for the lender in terms of both credit scorecard development and potential impact on existing customers.

Lenders use credit scoring models to assess potential borrowers' creditworthiness. To assign a credit score to an individual, these models typically use a variety of data, including credit history, employment status, income, and other factors. The credit score is then used to calculate the interest rate and other loan terms. Traditional credit scoring models, on the other hand, may not accurately capture the credit risk of certain individuals, especially those with lower credit scores. Lenders may consider using more and different data in their credit scoring models, as well as implementing risk-based pricing, to address this issue. We discuss the lender's implications for these strategies, with a focus on credit scorecard development and the potential impact on existing customers using More and Different Data in Credit Scoring Models

Traditional credit scoring models rely heavily on credit history data, such as payment history, credit utilization, and credit history length. However, these models may not fully capture an individual's credit risk, particularly for those with limited credit histories or who have previously faced financial difficulties. To address this issue, lenders may consider incorporating more and different data into their credit scoring models, such as alternative credit data or data from non-traditional sources.

Rent payments, utility bills, and other non-credit sources of payment history are examples of alternative credit data. Social media activity, education, and employment history are examples of non-traditional data sources. These types of data can provide additional insights into a person's creditworthiness, which is especially useful for those with limited credit histories or who may be underserved by traditional credit scoring models.

The use of alternative and non-traditional data sources in credit scoring models, on the other hand, raises several concerns. For starters, the accuracy and dependability of these data sources may be questionable. Second, using these data sources may introduce bias into the credit scoring model, especially if the data sources are not representative of the population under consideration. Finally, the use of non-traditional data sources may raise privacy concerns for those whose data is being used.

To address these concerns, lenders should carefully assess the quality and dependability of any additional data sources used in credit scoring models. They should also ensure that the use of these data sources is transparent and that individuals can understand and correct any errors in their data.

Risk-Based Pricing

Lenders use risk-based pricing to adjust the interest rate and other terms of a loan based on the borrower's credit risk. This approach enables lenders to make loans to people with low credit scores, but at a higher interest rate to reflect the increased risk of default.

Both lenders and borrowers can benefit from risk-based pricing. Lenders can expand their customer base and increase profitability by lending to people with lower credit scores. It gives borrowers access to credit that they might not have had otherwise.

However, risk-based pricing also raises several concerns. First, it may result in higher costs for individuals with lower credit scores, which can further exacerbate financial difficulties. Second, it may lead to discrimination against certain groups, such as low-income or minority individuals, who are more likely to have lower credit scores. Finally, it may lead to increased competition among lenders, which can result in aggressive marketing practices and a race to the bottom in terms of interest rates and other loan terms.

Lenders should ensure that risk-based pricing is applied fairly and transparently to address these concerns. They should also consider the potential impact on existing customers, especially those who may be facing financial difficulties. Furthermore, they should be open about the factors that influence loan terms and provide individuals with clear and concise information about the cost of credit.

Credit Scorecard Development

Developing a credit scorecard that incorporates more and different data sources and risk-based pricing. Let us compare the difference between Good and Bad Credit.

Bad Credit:

When lenders check your credit history, they may see some kinds of financial behavior as a red flag. If possible, you should avoid or minimize these to keep your score as high as possible:

- Frequently setting up new accounts. Opening a new bank account should only lower your credit score temporarily – but if you do it too often, your score won't have time to recover.
- Being close to your credit limit. Try not to max out your credit card or use your entire overdraft, as lenders may think you're over-reliant on credit or in financial difficulty.
- Applying for credit too often. Multiple credit applications can negatively affect your score, regardless of whether they're successful. This is because each application records a hard search on your report. Try to only apply for credit you're eligible for.
- Missing payments. If you miss a series of regular payments to lenders, they may record a default on your report. This can significantly lower your credit score for up to six years.
- Borrowing more than you can afford. If you can't pay off your debts, you may have to get a Debt Relief Order or Individual Voluntary Arrangement. Lenders can also try to reclaim money you owe by getting a County court judgment (such as a County Court Judgment) issued against you, or by applying to make you bankrupt. Any of these events will significantly reduce your credit score and make it difficult to borrow money or even open a bank account in the future.
- Having little or no credit history. If you've never had credit you'll likely to have a low credit score. This is because lenders like to see a good track record of sensible borrowing, which helps them decide if you're likely to pay them back on time.

Good Credit:

Building a good credit score takes time and consistent effort. Here are some steps we can take to build and maintain a good credit score:

1. Pay your bills on time: Late payments can have a negative impact on your credit score, so make sure you pay all your bills on time.
2. Keep your credit utilization low: Your credit utilization is the amount of credit you use compared to your credit limit. Try to keep your credit utilization below 30% to build a good credit score.
3. Open credit accounts carefully: Opening too many credit accounts at once can hurt your credit score. Start with a few credit accounts that you can manage responsibly.
4. Monitor your credit report: Check your credit report regularly to make sure all the information is accurate. If you find any errors, dispute them with the credit bureau.
5. Use different types of credit: Having a mix of different types of credit, such as credit cards, installment loans, and mortgages, can help build a strong credit score.
6. Keep old credit accounts open: Length of credit history is an important factor in determining your credit score. Keep old credit accounts open, even if you don't use them frequently.

Giving credit to bad customers can have several implications for lenders. Here are some of the potential impacts:

1. Increased credit risk: Bad customers are more likely to default on their loans or miss payments, which increases the lender's credit risk. This can result in higher losses and lower profitability for the lender.
2. Lower credit scorecard performance: If bad customers are included in the credit scorecard development, it can negatively impact the performance of the scorecard. The scorecard may not accurately predict creditworthiness, leading to more defaults and higher losses.
3. Higher interest rates: Lenders may need to charge higher interest rates to bad customers to compensate for the increased credit risk. This can make it more difficult for these customers to repay their loans and may lead to more defaults.
4. Negative impact on existing customers: If lenders give credit to bad customers and experience high levels of defaults, it can have a negative impact on existing customers. The lender may need to tighten credit standards, resulting in fewer loan approvals and less credit availability for existing customers making customers move to other banks.
5. Risk of Reputation: If the lender is known for giving credit to bad customers and experiencing high levels of defaults, it can harm their reputation. This can lead to a decrease in business and difficulty attracting new customers.

Thus, Lenders face many issues during the development of scorecard and has several impacts on customers.

Part B:

The 'GermanCreditData' dataset that we are examining here has 1000 records with 22 fields and includes credit information. The classification's objective is to predict whether the consumer will ultimately be a Good or Bad Credit Risk.

The target value is indicated as either "Bad" or "Good" in the various categories that carry credit information about the clients shown in the preceding figure. We removed the field "Bad" from the data set because these two values are directly opposed to one another. We search for missing data as we continue to explore the data. According to the data dictionary, several of the fields' data types need to be modified because their category features come in the form of numerals.

We can deduce the following by looking at the mean value of the numerical variables when we group them by the target output, "Good":

- It makes sense that a good consumer takes little time.
- Good customers take out less credit than bad customers do.
- Understandably, good customers also have a cheap instalment rate.
- The target value is not significantly impacted by the length of residence.
- It does seem reasonable that good customers are older.
- The results are not severely affected by the current credits.

```
(1000, 22)
Index(['Checking', 'Duration', 'History', 'Purpose', 'Amount', 'Savings',
       'Employed', 'Installp', 'marital', 'Coapp', 'Resident', 'Property',
       'Age', 'Other', 'housing', 'Existcr', 'Job', 'Depends', 'Telephone',
       'Foreign', 'Bad', 'Good'],
      dtype='object')
```

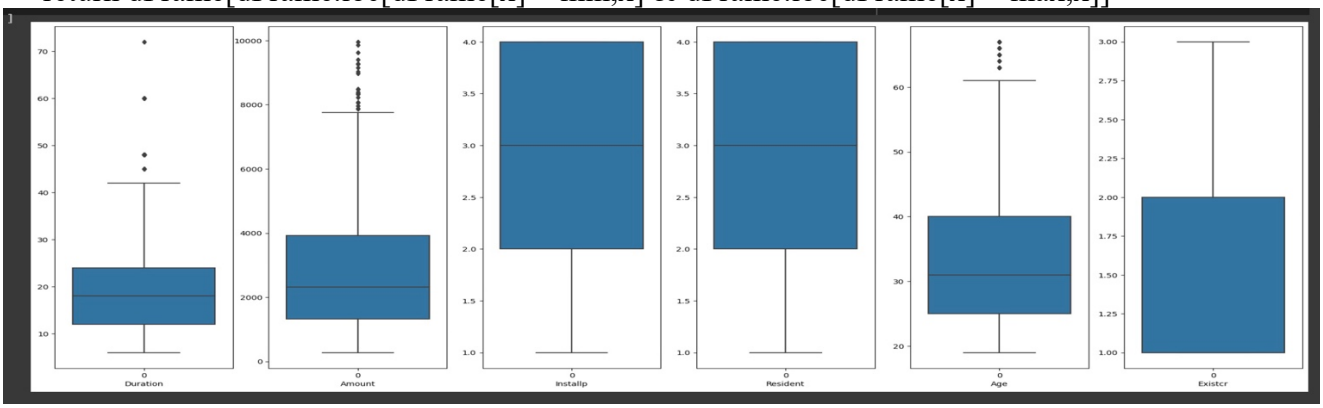
Task 1: Splitting the dataset into two subsets:

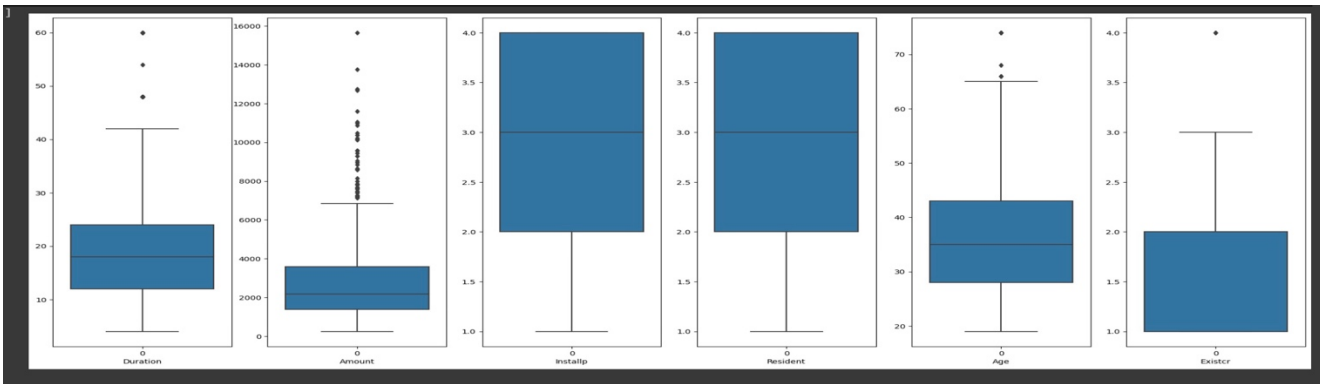
We must split it based on Checking, subset 1 is (checking==1 and checking==2) Subset 2 is (checking==3 and checking==4).We then clean the subsets for necessary outliers. We delete all outliers which are not present in the interquartile range. We can visualize it using boxplots and notice multiple outliers for various columns.

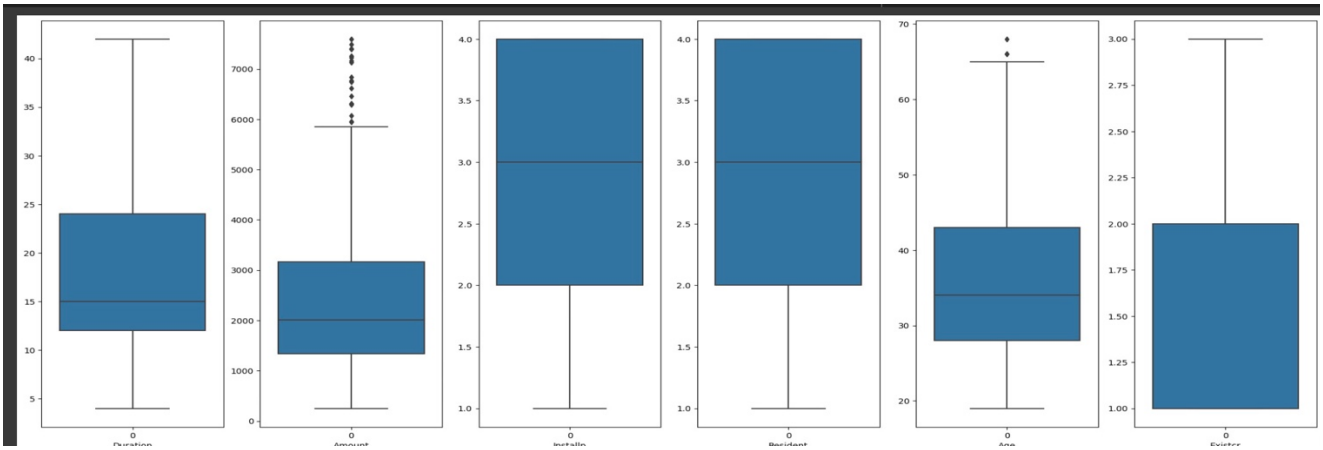
Code snippet:

```
delete_outlier: function to delete outliers
args: num_col - numerical column whose outliers needs to be deleted
     dFrame - data frame from which the column is picked
returns: data frame after deleting the outliers
def delete_outlier(num_col, dFrame):
  for x in [num_col]:
```

```
q75,q25 = np.percentile(dFrame.loc[:,x],[75,25])
intr_qr = q75-q25
max = q75+(2*intr_qr)
min = q25-(2*intr_qr)
 return dFrame[dFrame.loc[dFrame[x] > min,x] & dFrame.loc[dFrame[x] < max,x]]
```







Task 2: Establishing a training and a validation set for each subset:

Validation is done to make sure the developed model is suitable to the target population and that it hasn't been overfit or underfit.

The Pareto Principle is also known as the 80/20 rule. The basic idea is that in most cases, 80% of effects result from 20% of causes. In terms of statistics, it comes close to explaining a wide range of environmental, mechanical, and human events. As a result, I divided my training and testing into an 80/20 split.

Consumer behavior changes over time, and lenders' targeting, underwriting, and collection methods are also constantly evolving. As a result, a model developed using data from one time period may not work effectively in another. We want to develop a model that employs traits and characteristics that tend to hold up well over time, regardless of changes in the lending environment (Mays and Niall Lynas 2010). Training and validation sets are frequently used to create credit risk models. To predict results for specific data, one must train a collection of data, which necessitates the use of a training set. A prediction that has not been validated, on the other hand, lacks credibility. As a result, a data set is always divided. Thus, I divided besting on training and testing set also called Validation. Thus no significant issues were faced.

Code snippet:

```
X_train2, X_val2, y_train2, y_val2 = train_test_split(X_sub2, y_sub2,
                                    test_size=0.20,
                                    stratify = y_sub2,
                                    random_state=42)
```

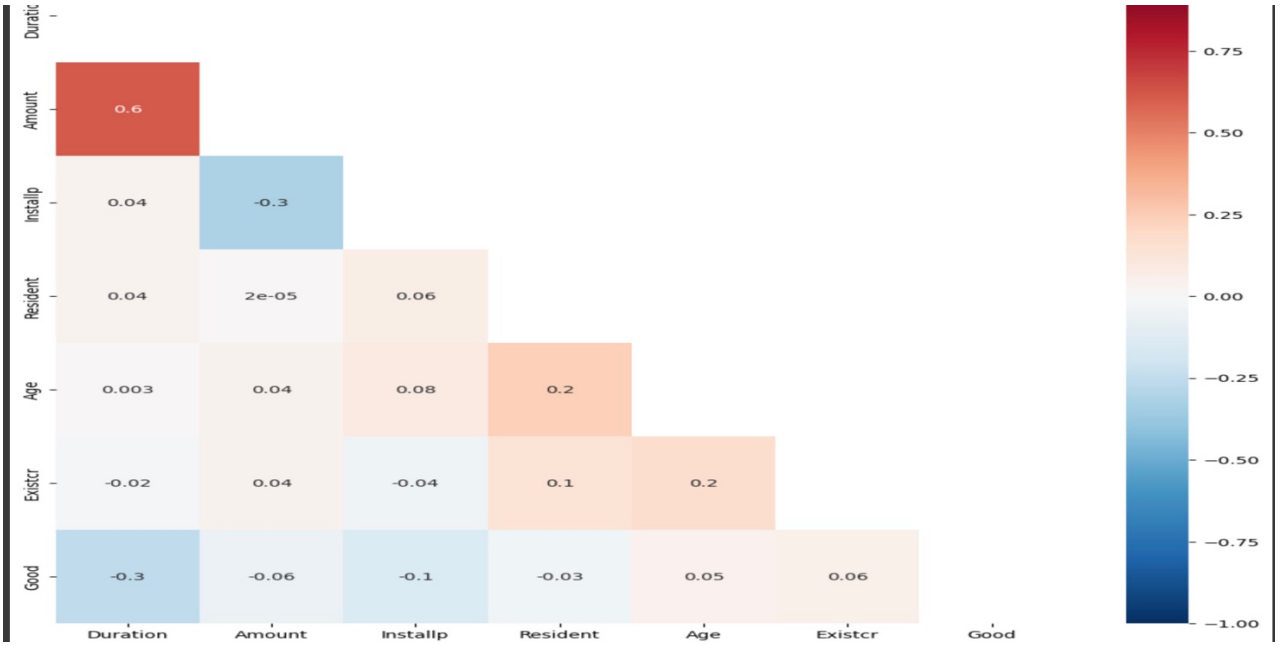
Task 3: Choosing four variables:

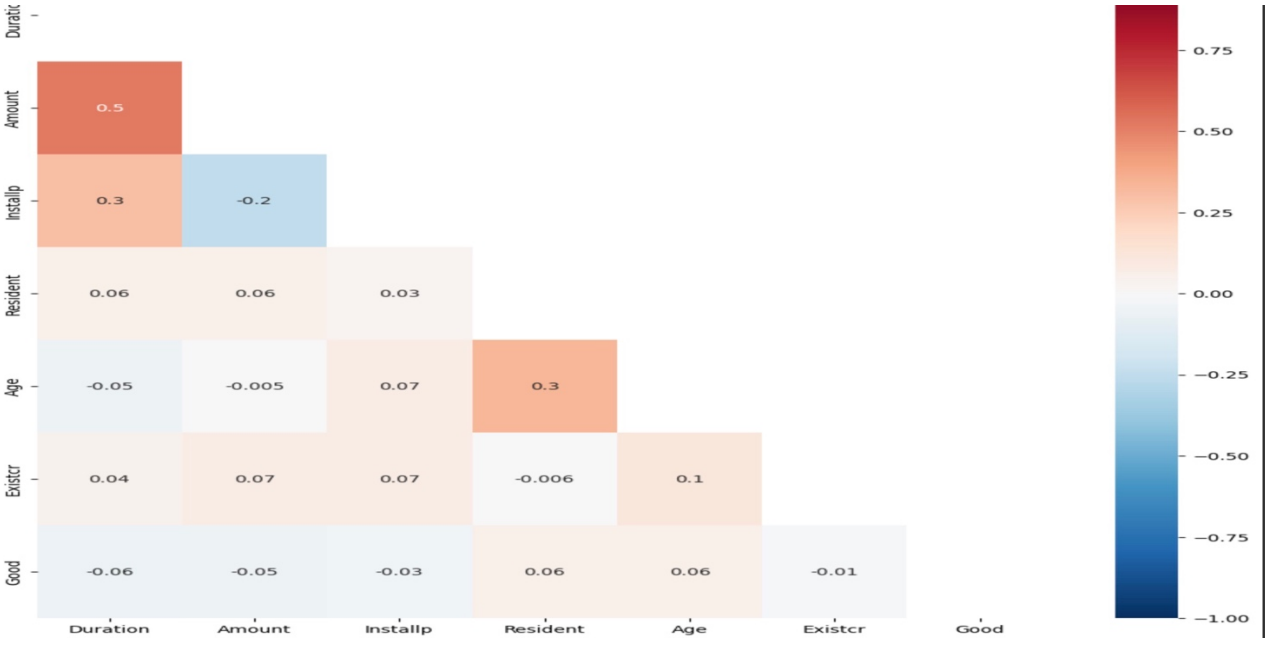
For each training set choose four variables which are suitable for building a scorecard. For each training set the variables must have (i) at least one continuous variable before binning; (ii) at least one categorical variable with more than two categories, so you can see whether categories can be combined.

We develop a heat map which displays the relationship between each integer variable and every other variable of the same sort. We see the correlation between the variables. For subset1, we choose Savings, Duration, Installp and History. Duration and Installp have the highest relative association with ‘Good’. For subset2, we choose Purpose, Duration, Emploed and History. Again, I choose Duration and Installp because they have highest relative association with good.

For categorical features selection, I utilized Chi-square test and picked top 3 categories with highest value. For subset 1, the categorical features with the highest chi-square values are Savings, History, and Property; for set 2, the features with the highest chi-square values are Purpose, History, and Emploed. I used regression iteratively over every potential combination of four features to choose the combination with the highest value, which gave me the four required

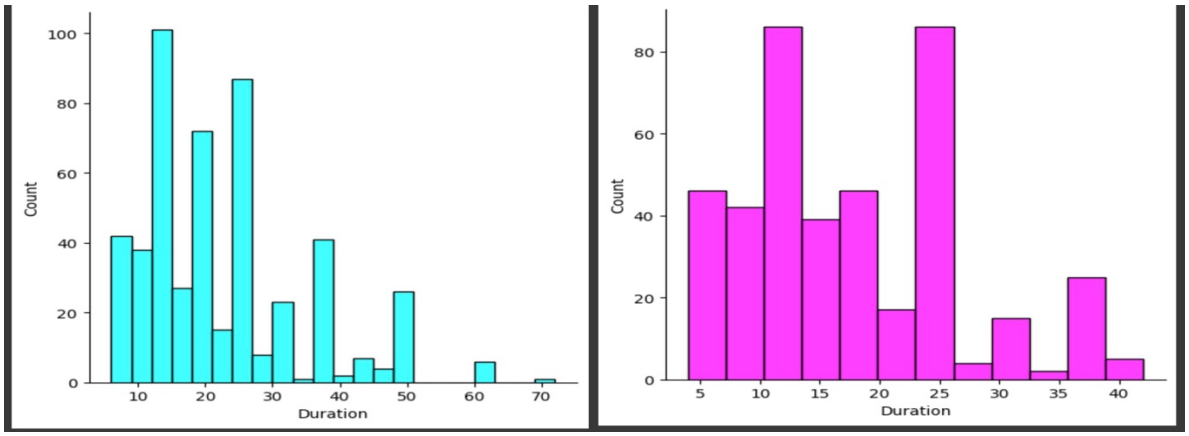
characteristics. I followed the same steps to acquire 4 attributes for the second subset. Using the procedure, I selected the features Savings, Duration, History, and Installp for subset 1. Additionally, the features Purpose, Duration, History, and Emploed were chosen from subset 2.





Task 4: Making of Score card:

Since the default risk is non-linear in a continuous variable, the characteristics must be transformed into binary variables before the scorecard can be built. Binding the numerical variables will allow us to transform the numerical characteristics into categorical ones, which is what we need to do. Both sets in this project share the same single numerical variable, Duration.



The distribution of both sets is shown in the histogram images. The bins are separated into two categories for the first set, which are [5, 18]and [18, 72]. The second set is divided into three categories which are [3, 12], [12,22]and [22, 42].

Using the get dummies function in pandas, the variables are encoded and converted into binary variables after being binned. Now, models are applied to the prepared data to train them. I have chosen linear regression and logistic regression, two of the basic models. The Appendix A below shows the binary variables utilized in these models along with their coefficients.

Logistic Regression subset1:

Confusion Matrix:
array ([[21, 23],
[13, 44]])
Accuracy Score:0.6435643564356436

Logistic Regression subset2:

Confusion Matrix:
array ([[0, 10],
[0, 73]])
Accuracy score: 0.8795180722891566

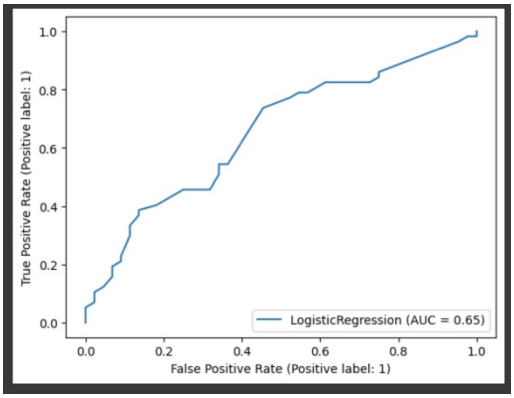
Linear Regression subset 1:

Confusion Matrix
[[24 47]
[15 74]]
Accuracy score:0.676470588

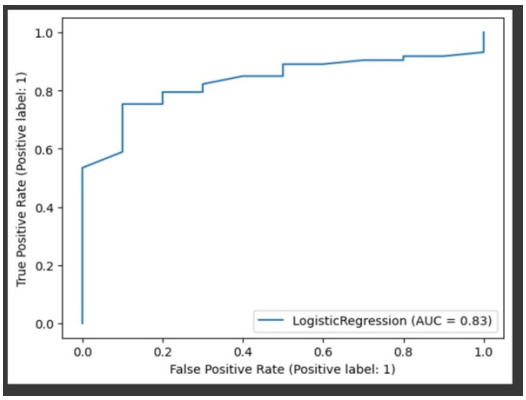
The confusion matrix and accuracy for the sets for logistic and linear regressions are shown. One may simply conclude that logistic regression performed better on both subsets by looking at accuracy in the figure above.

Task 5: ROC Curve:

Logistic Regression set 1



Logistic Regression set 2



After examining the ROC curve for the logistic regression subset 1 and subset 2, we notice subset 2's curves are superior and better and that the second set of the linear regression has the best output as it has the largest area under its cover.

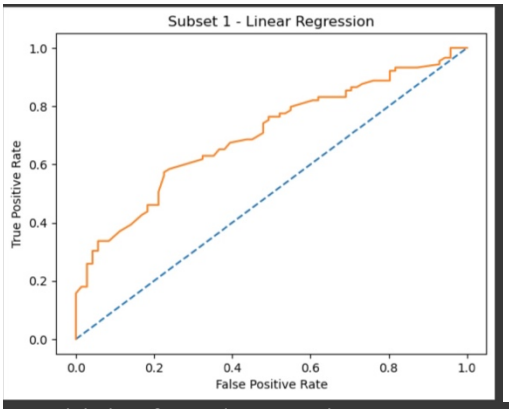
Logistic Regression set 1

Sensitivity for subset 1: 0.6176470588235294
Specificity for subset 1: 0.6567164179104478
Area under the ROC curve of subset 1 is: 0.6246012759
Gini coefficient for subset 1: 0.24920255183413076
KS statistics value for subset 1 is: 0.2743634767339772

Logistic Regression set 2:

Sensitivity for subset 2: nan
Specificity for subset 2: 0.8795180722891566
Area under the ROC curve of subset 2 is: 0.5
Gini coefficient for subset 2: 0.0
KS statistics value for subset 2 is: nan

Linear Regression



Sensitivity for subset 1 Linear: 0.6153846153846154
Specificity for subset 1 Linear: 0.6115702479338843
Gini coefficient for subset 1 Linear: 0.39927203671467004
KS statistics value for subset 1 Linear is: 0.22695486331849962

Like ROC curve, the evaluation parameters also shows that subgroup 2 has a superior result, and linear regression has performed better in this regard. We can see that the second set's linear regression has the greatest area under the ROC curve, or if we consider the better Gini or KS value. In fact, it is the sole value that is higher than 50%, which is great as it shows that the model is efficient in more successfully identifying good customers.

Identifying good and bad consumers was the aim of the modelling process. It is obvious that in the given scenario, identifying a bad customer is more important. As a result, we will concentrate on the specificity score, which will inform us of the true negative value. By comparing the specificity scores for the two subsets, we can state with certainty that the models will be able to identify bad customers with ease. However, the values of the KS statistics show that this model needs additional improvement as they are low and show that it is decent at correctly differentiating between defaulter and non-defaulter.

Appendix A

The Binary Variables and their coefficients used in subset 1 of Linear Regression:

['Savings_2', 'Savings_4', 'Savings_5', 'Duration_bin(18, 72]', 'History_1', 'History_2', 'History_3', 'History_4', 'Installp_2', 'Installp_3', 'Installp_4']

[0.08478952 0.3251005 0.18775665 -0.17938461 0.00294928 0.18974043 0.30035811 0.314347 -0.0319885 -0.14605395 -0.20374096]

The Binary variables and their coefficients used in subset 2 of Linear Regression:

['Purpose_1', 'Purpose_2', 'Purpose_3', 'Purpose_4', 'Purpose_6', 'Purpose_8', 'Purpose_9', 'Duration_bin_(12, 22]', 'Duration_bin_(22, 42]', 'History_1', 'History_2', 'History_3', 'History_4', 'Emploed_2', 'Emploed_3', 'Emploed_4', 'Emploed_5']

[0.14072563 0.03580755 0.05777858 0.20211837 -0.12440055 0.24965958 -0.01872114 -0.066396 -0.03796859 0.0666191 0.28622642 0.25752405 0.35752778 0.06506836 0.12034375 0.1407244 0.16980883]

The Binary variables and their coefficients used in subset 1 of Logistic Regression:

['Savings_2', 'Savings_4', 'Savings_5', 'Duration_bin_(18, 72]', 'History_1', 'History_2', 'History_3', 'History_4', 'Installp_2', 'Installp_3', 'Installp_4']

[0.29054068 1.21758182 0.80826005 -0.76640586 -0.29612737 0.47930063 0.88085366 1.01413182 -0.05436707 -0.51773297 -0.7826345]

The Binary Variable and their coefficients used in subset 2 of Logistic Regression:

['Purpose_1', 'Purpose_2', 'Purpose_3', 'Purpose_4', 'Purpose_6', 'Purpose_8', 'Purpose_9', 'Duration_bin_(12, 22]', 'Duration_bin_(22, 42]', 'History_1', 'History_2', 'History_3', 'History_4', 'Emploed_2', 'Emploed_3', 'Emploed_4', 'Emploed_5']

[1.09375982 0.20805243 0.44395061 0.43878034 -0.68324477 0.45050753 -0.17431798 -0.44467034 -0.10693999 -0.38841852 0.68824755 0.36137121 1.34790085 -0.08333834 0.29131525 0.51986442 0.76507197]

Reference List:

Anukrati Mehta 2019. A Beginner's Guide to Credit Risk Modelling. Available at: https://www.digitalvidya.com/blog/credit-risk-modelling/.

Credit Scoring and Financial Inclusion - research study. 2021. Available at: https://vantagescore.com/credit-scoring-and-financial-inclusion-research-study/ [Accessed: 1 April 2022].

What Factors Determine Credit Risk? Available at: https://www.investopedia.com/ask/answers/022415/what-factors-are-taken-account-quantify-credit-risk.asp.

Siddiqi, N. 2006. Credit risk scorecards: developing and implementing intelligent credit scoring. Hoboken, N.J.: Wiley.

Thomas, L.C. 2000. A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting 16(2), pp. 149–172. doi: 10.1016/s0169-2070(00)00034-0.

Thomas, L.C., Oliver, R.W. and Hand, D.J. 2005. A survey of the issues in consumer credit modelling research. Journal of the Operational Research Society 56(9), pp. 1006–1015. doi: 10.1057/palgrave.jors.2602018.