Credit Risk Analysis

By-

- Raj Shah (202218029)
- Shivani Pande (202218044)
- Swapnil Sheth (202218045)
- Vipul Singh (202218052)
- Jitul Bakshi (202218059)

What is Credit Risk Analysis?

- Credit risk analysis is the process of assessing the potential risk involved in lending money to individuals or entities.
- It involves gathering relevant data, evaluating creditworthiness, and estimating the likelihood of default or other credit-related issues.
- By analyzing factors such as credit history, financial stability, and repayment capacity, credit risk analysis helps lenders make informed decisions about lending, setting interest rates, and managing risk exposure. It involves credit scoring, financial analysis, risk modeling, and portfolio analysis to evaluate the overall credit risk and make sound credit decisions.

Our purpose

To accurately assess the creditworthiness of borrowers and predict the likelihood of default or other credit-related events based on their financial information and credit history. This helps lenders make informed decisions regarding loan approvals, interest rates, and risk management strategies.

Build a strategy to shortlist customers who can be offered a fresh loan.

01

Data set Description

For the purpose of the project we were using German credit card data set.(1000 * 10)

Name: Sex
Male: 690 Female :310

Name: Job
skilled 630 unskilled_and_non-resident 222 highly skilled 148

Name: Housing
own 713 rent 179 free 108

Name: Saving accounts
little 786, moderate 103 quite rich 63, rich 48

Name: Checking account
moderate 472 little 465 rich 63

Name: Purpose

car	337
radio/TV	280
furniture/equipment	181
business	97
education	59
repairs	22
vacation/others	12
domestic appliances	12

Target Variable

Name: Risk

Good: 700
Bad: 300

We have added another column in our dataset - Job Category, which is made using the Job column where we have mapped unique category of Jobs as follows -

0 - Unskilled and Non-resident

1 - Unskilled and Resident

2- Skilled

3- Highly Skilled

Job	Job Category
2	Skilled
2	Skilled
1	Unskilled and Resident
2	Skilled
2	Skilled

#	Column	Non-Null Count	Dtype
0	Age	1000 non-null	int64
1	Sex	1000 non-null	object
2	Job	1000 non-null	int64
3	Housing	1000 non-null	object
4	Saving accounts	817 non-null	object
5	Checking account	606 non-null	object
6	Credit amount	1000 non-null	int64
7	Duration	1000 non-null	int64
8	Purpose	1000 non-null	object
9	Risk	1000 non-null	object
10	Job Category	1000 non-null	object

- There are total 10 columns and 1,000 observations in the dataset.
- We have only three continuous variables - Age, Credit Amount, and Duration.
- All other variables are categorical.
- There is missing values in the dataset. By dropping the rows with no values, the dataset will lose 577 instances which is more than half.
- That’s a significant loss of data.
- To avoid this, we replaced the null values with none. It’s possible that the applicants with null values didn’t have a savings or checking account at the time of the application.

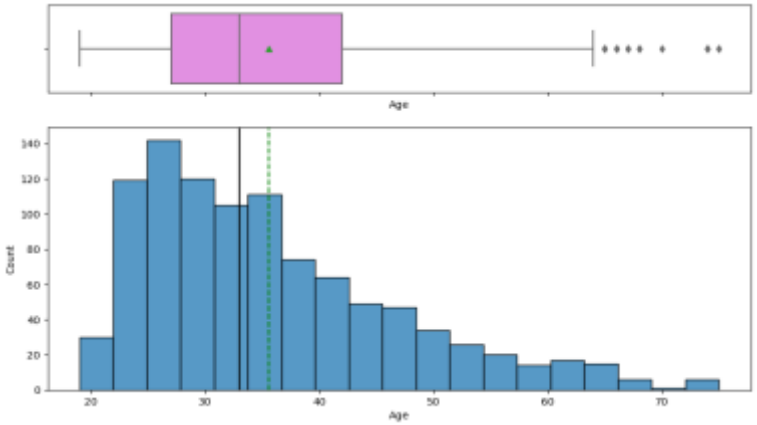
	count	mean	std	min	25%	50%	75%	max
Age	1000.0	35.546	11.375469	19.0	27.0	33.0	42.00	75.0
Job	1000.0	1.904	0.653614	0.0	2.0	2.0	2.00	3.0
Credit amount	1000.0	3271.258	2822.736876	250.0	1365.5	2319.5	3972.25	18424.0
Duration	1000.0	20.903	12.058814	4.0	12.0	18.0	24.00	72.0

- Mean value for the age column is approx 35 and the median is 33. This shows that majority of the customers are under 35 years of age.
- Mean amount of credit is approx 3,271 but it has a wide range with values from 250 to 18,424. We will explore this further in univariate analysis.
- Mean duration for which the credit is given is approx 21 months.

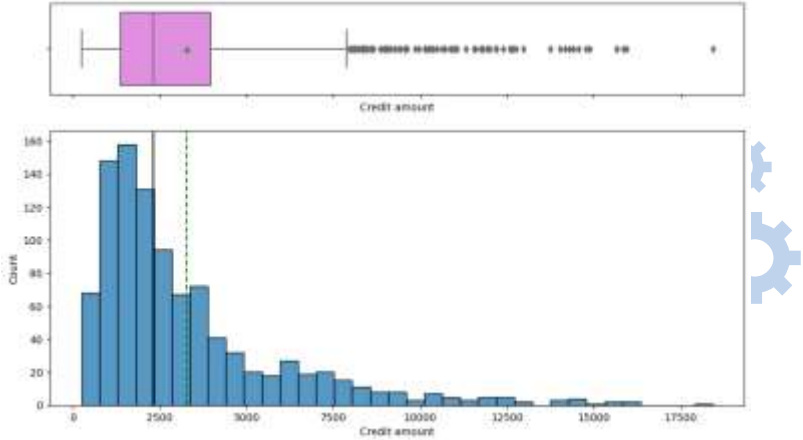
02

Exploratory Data Analysis

2.1
Univariate Analysis



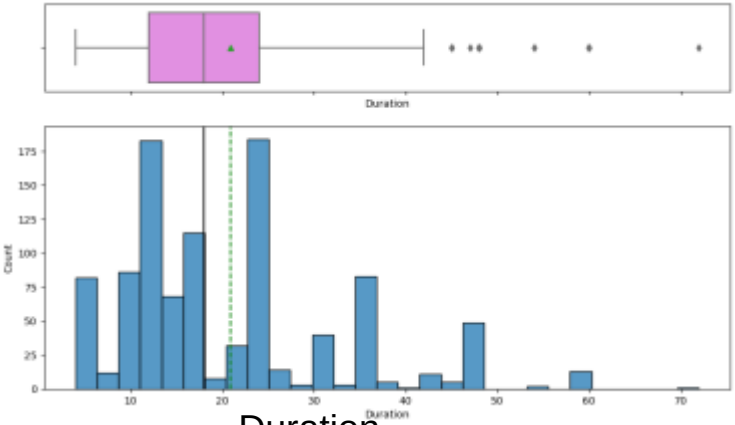
Age



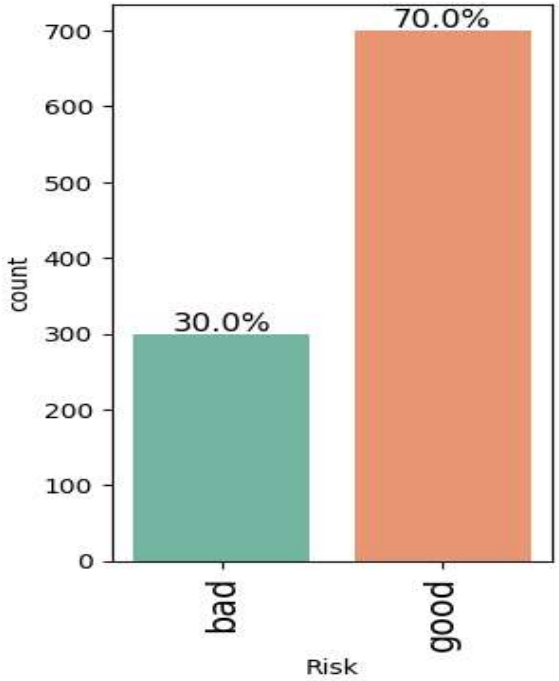
Credit amount

Observation:

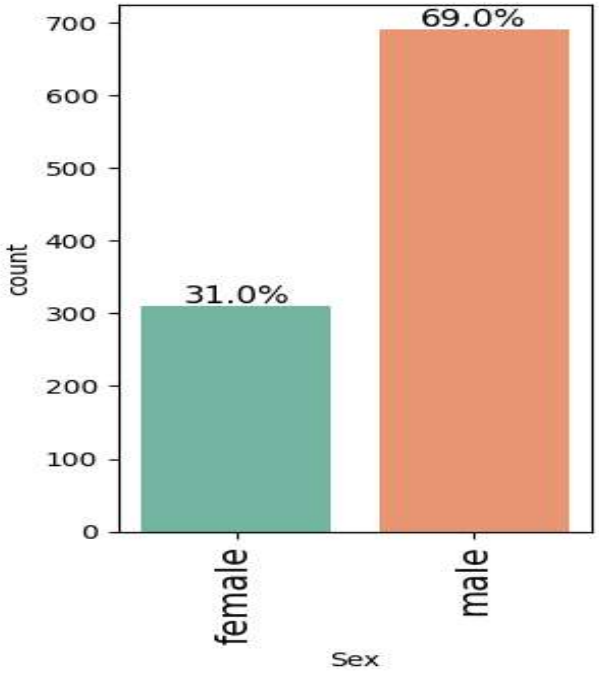
- The distribution of Age, Credit Amount and Duration is right-skewed.
- The boxplot shows that there are outliers at the right end.
- We will not treat these outliers as they represent the real market trend.



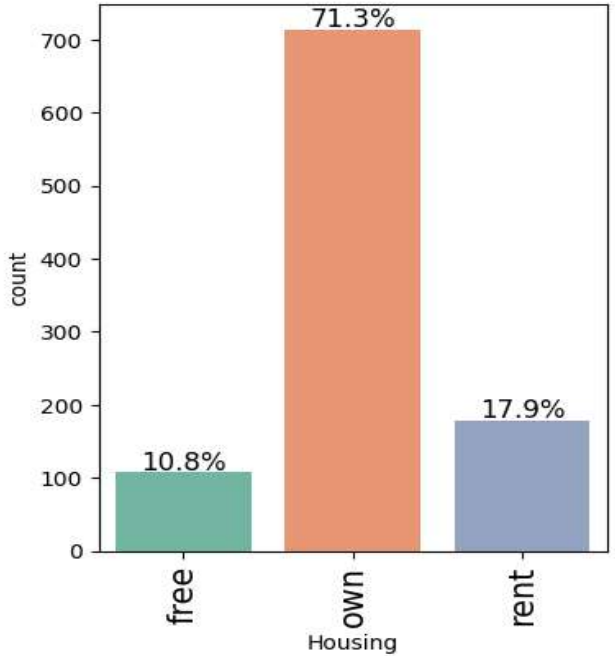
Duration



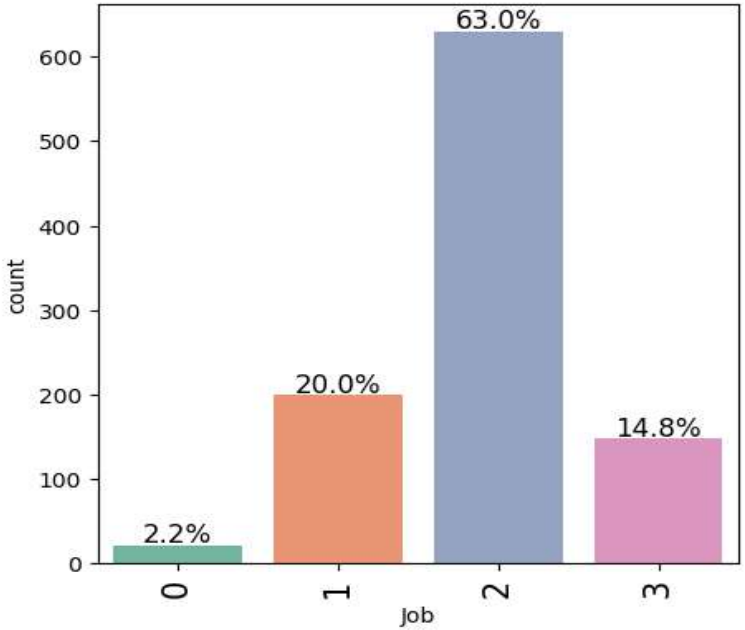
- The class distribution in the target variable is imbalanced.
- We have 70% observations for non-defaulters and 30% observations for defaulters.



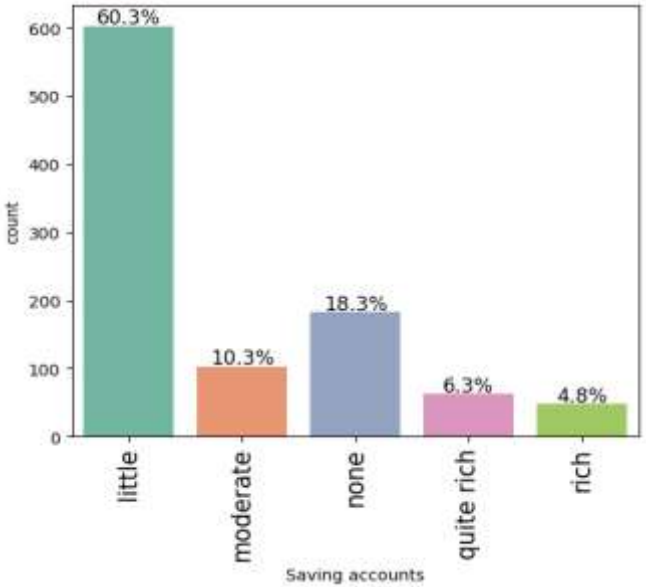
- Male customers are taking more credit than female customers
- There are 69% male customers and 31% female customers



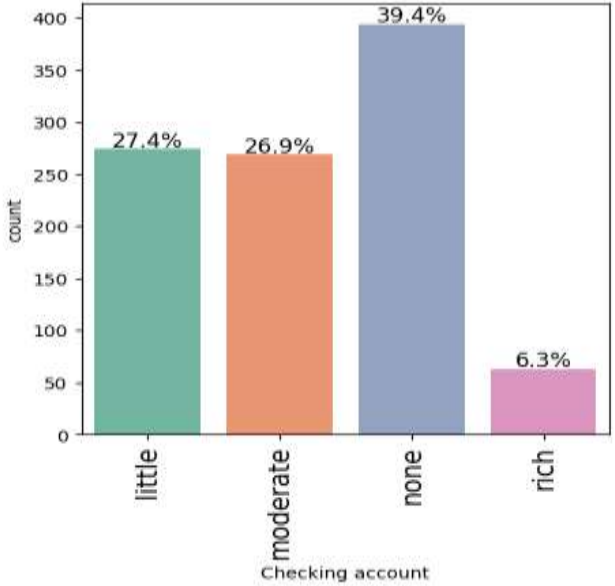
- Major of the customers, approx 71%, who take credit have their own house
- Approx 18% customers are living in a rented house.
- There are only 11% customers who have free housing. These are the customers who live in a house given by their company or organization



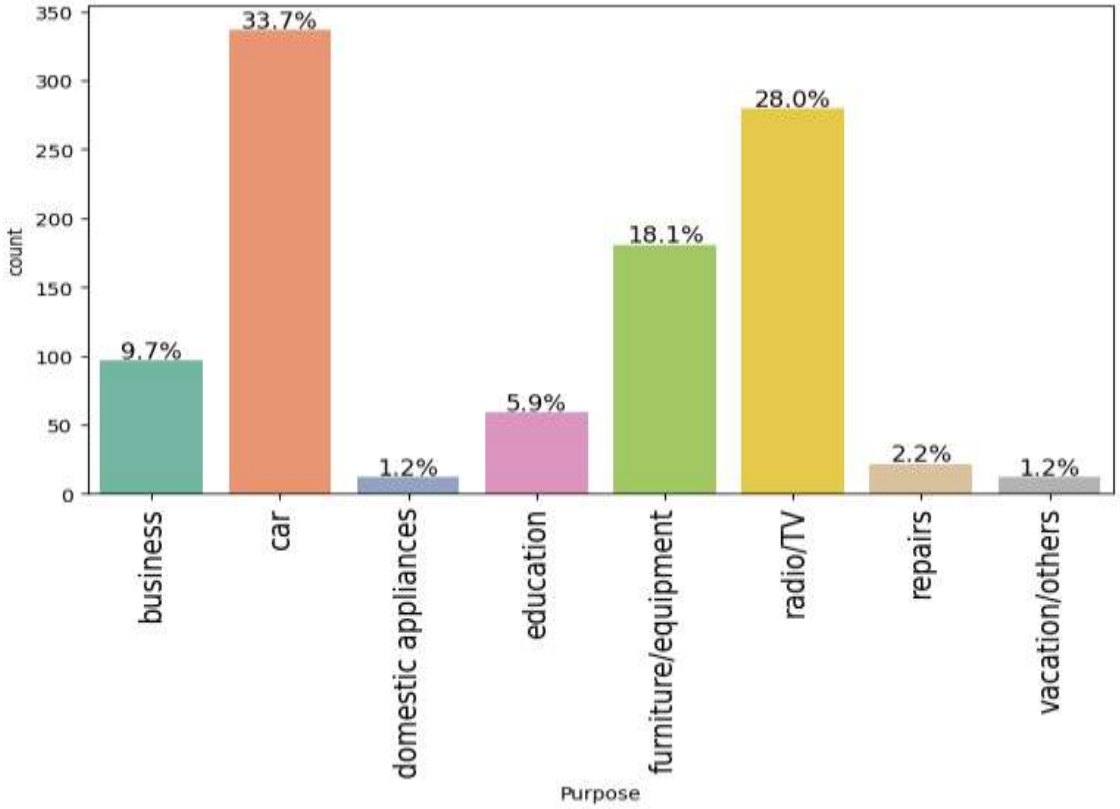
- Majority of the customers i.e. 63% fall into the skilled category.
- There are only approx 15% customers that lie in highly skilled category which makes sense as these may be the persons with high education or highly experienced.
- There are very few observations, approx 22%, with 0 or 1 job category.



- Approx 70% customers who take credit have a little or moderate amount in their savings account. This makes sense as these customers would need credit more than the other categories.
- Approx 11% customers who take credit are in the rich category based on their balance in the savings account.
- Note that the percentages do not add up to 100 as we have missing values in this column.



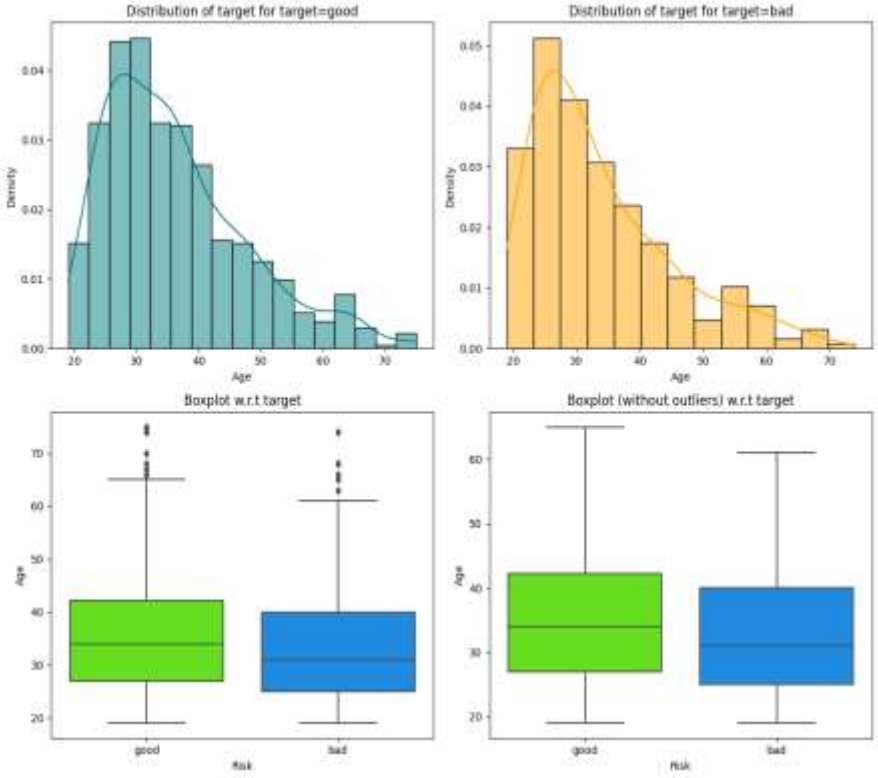
- Approx 54% customers who take credit have a little or moderate amount in their checking account. This makes sense as these customers would need credit more than the other categories.
- Approx 6% customers who take credit are in the rich category based on their balance in the checking account.
- Note that the percentages do not add up to 100 as we have missing values in this column.



- The plot shows that most customers take credit for luxury items like car, radio or furniture/equipment, domestic appliances.
- Approximately just 16% customers take credit for business or education

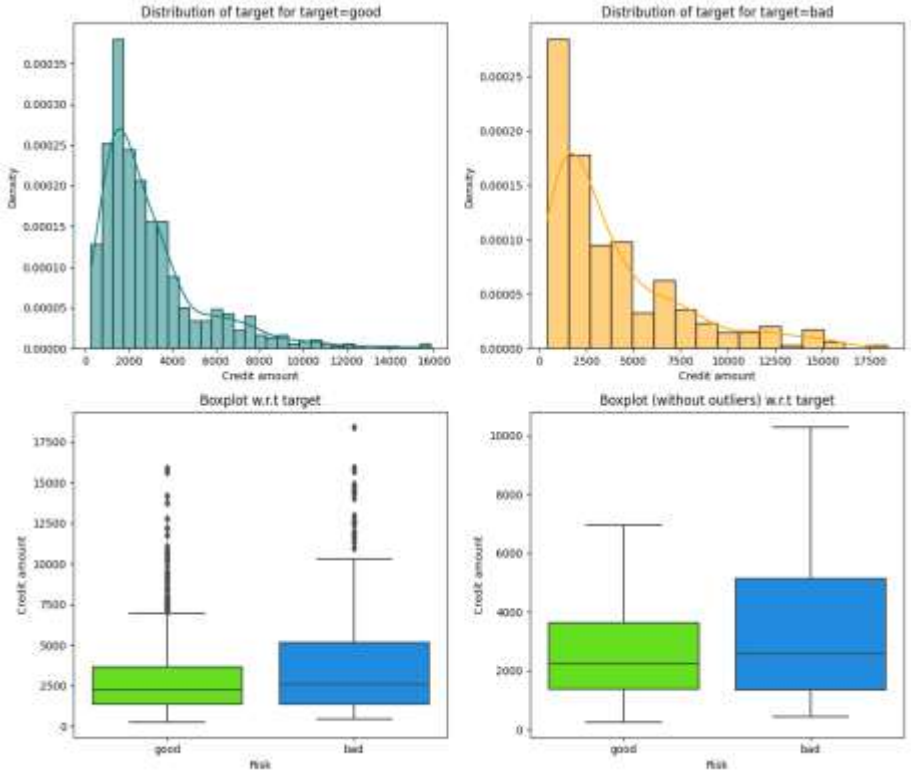
2.2
Bivariate Analysis

Risk vs Age



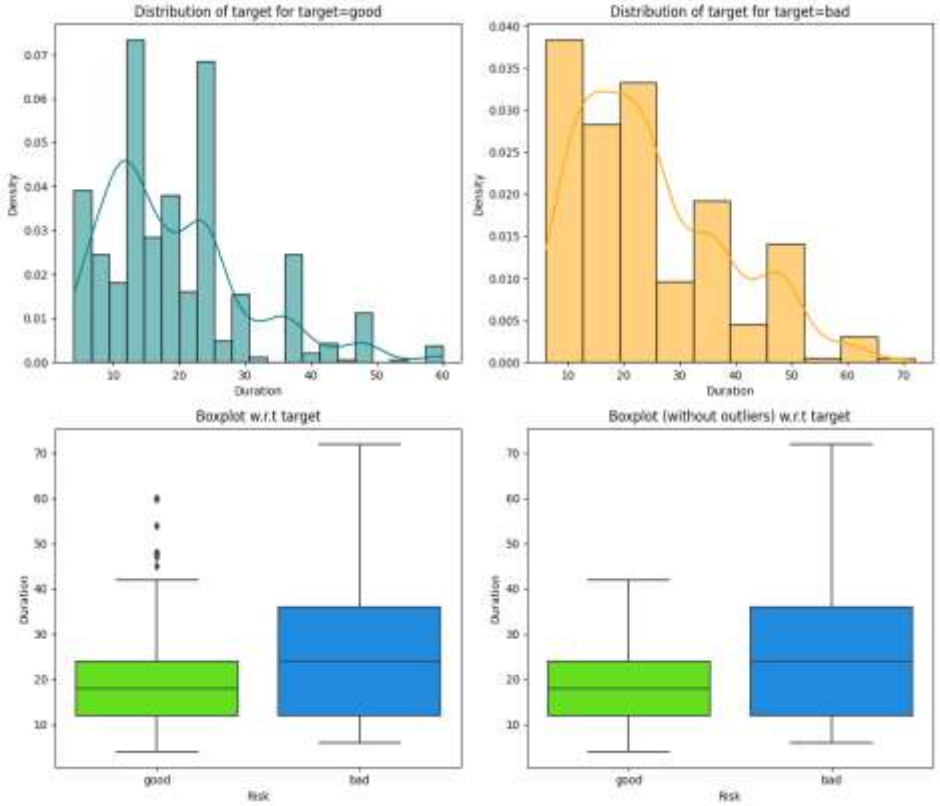
- We can see that the median age of defaulters is less than the median age of non-defaulters.
- This shows that younger customers are more likely to default.
- There are outliers in boxplots of both class distributions. These outliers are mainly in the elderly range. But the density of the customer is also low in comparison to other age groups.

Risk vs Credit Amount

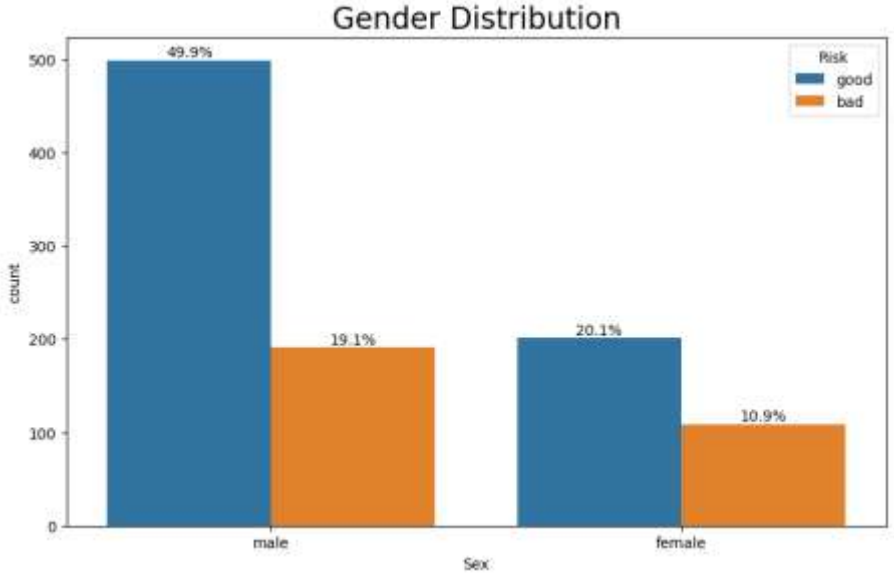


- We can see that the third quartile amount of defaulters is much more than the third quartile amount of non-defaulters.This shows that customers with high credit amount are more likely to default.
- The distribution of Credit Amount is heavily right skewed for both Good and Bad loan.
- There are outliers in boxplots of both class distributions.

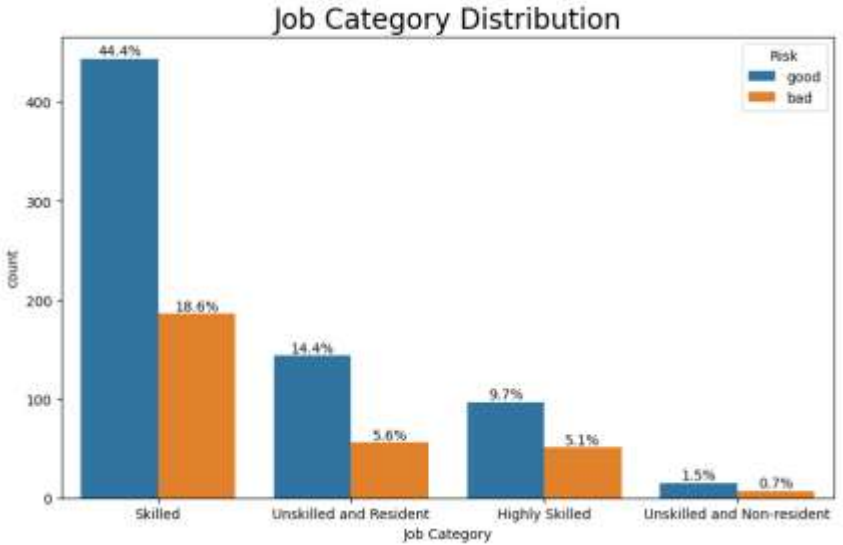
Risk vs Duration



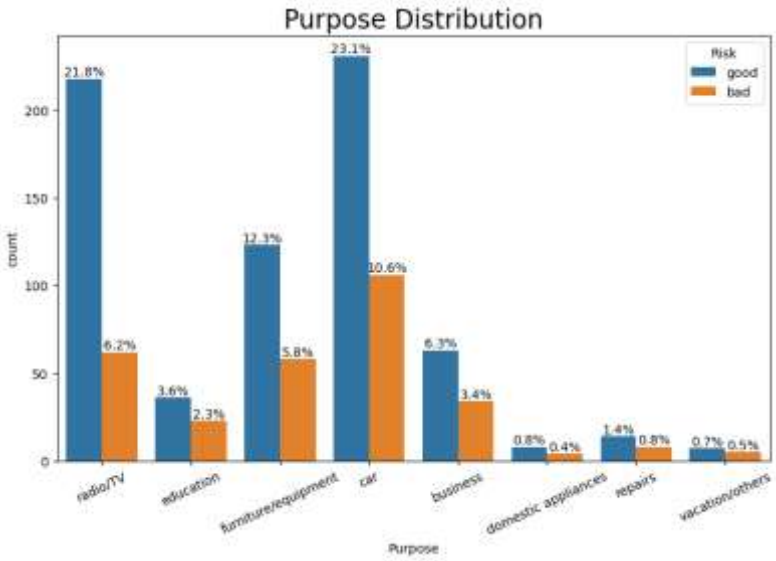
- We can see that the second and third quartiles of duration of defaulters is much more than the second and third quartiles duration of non-defaulters.
- This shows that customers with high duration are more likely to default.

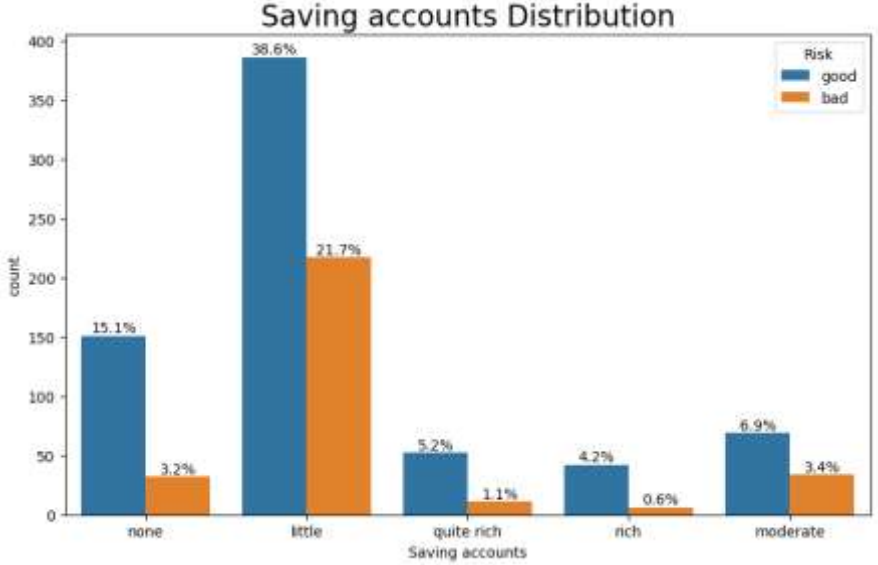


- We saw earlier that the percentage of male customers is more than the female customers. This plot shows that female customers are more likely to default as compared to male customers.
- About 2/5 of male applicants and 1/3 of female applicants are classified as bad.



- More than 50% of applicants are under the skilled and unskilled and resident job categories.
- Applicants that are highly skilled are more likely to take out larger loans.
- Highly skilled or unskilled non-resident customers are more likely to default as compared to skilled and unskilled and resident customers.



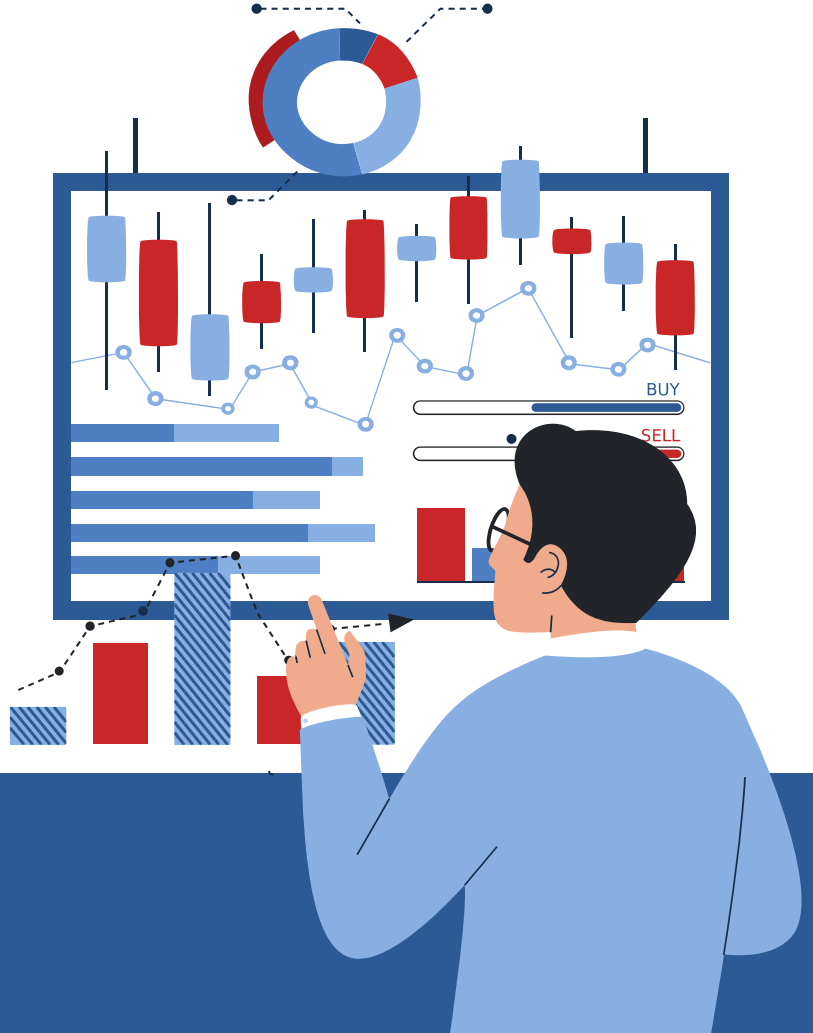


- Customers who take credit for radio/TV are least likely to default. This might be because their credit amount is small.
- Customers who take credit for education or vacation are most likely to default.
- Other categories have no significant difference between their default and non-default ratio.

- As we saw earlier, customers with little or moderate amounts in saving accounts takes more credit but at the same time they are most likely to default.
- Rich customers are slightly less likely to default as compared to quite rich customers.

03

Machine Learning



In order to construct the model, it is essential to determine the nature of the problem. Is it a supervised learning problem, an unsupervised learning problem, or another type altogether? Furthermore, is it a classification task, a regression task, or something different entirely? Through our exploratory data analysis, we have gathered that the data consists of labeled target variables, where each instance is accompanied by an expected output. Additionally, we have identified this as a classification task since the objective is to predict a specific class. To be more precise, it is a binary classification task since the target variable consists of two distinct types:

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	67	male	skilled	own	NaN	little	1169	6	radio/TV	good
1	22	female	skilled	own	little	moderate	5951	48	radio/TV	bad
2	49	male	unskilled and resident	own	little	NaN	2096	12	education	good
3	45	male	skilled	free	little	little	7882	42	furniture/equipment	good
4	53	male	skilled	free	little	little	4870	24	car	bad

Before constructing the model, it is necessary to perform data transformation through a technique called feature engineering. This process involves manipulating the dataset by adding, deleting, combining, and other operations to enhance the machine learning model's performance.
During the exploratory data analysis, we have identified the presence of missing values in the dataset. Therefore, the initial transformation step will involve imputing instances that contain missing values with None.

```
Age               False
Sex               False
Job               False
Housing           False
Saving accounts   False
Checking account  False
Credit amount     False
Duration          False
Purpose           False
Risk              False
dtype: bool
```

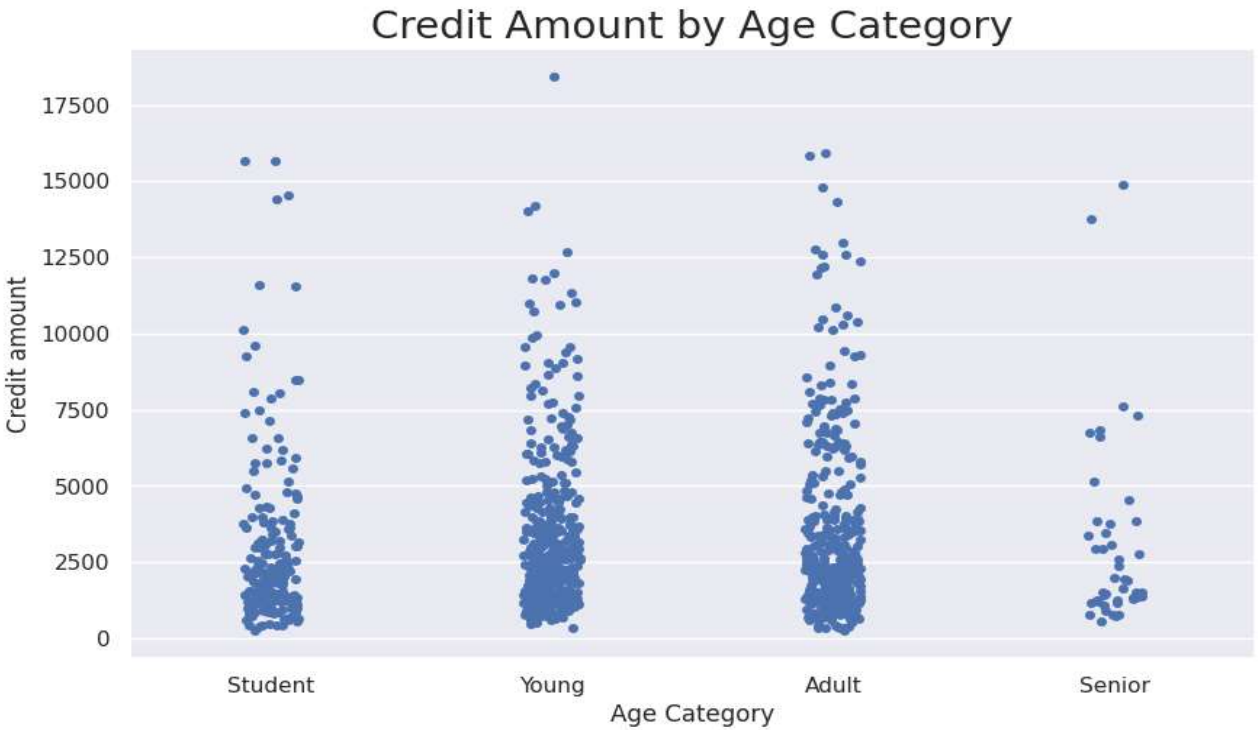
Due to the insufficient information contained within the features of the dataset, it is necessary to generate new features. In light of this, we have created two additional features, namely "Age Category" and "Credit Category."

The "Age Category" feature is designed to categorize individuals based on their age range. Specifically, it includes the following categories: "Student" (18-24 years), "Young" (25-34 years), "Adult" (35-59 years), and "Senior" (60-119 years).

Similarly, the "Credit Category" feature follows a similar approach to categorize individuals based on their credit status. The specific categories within the "Credit Category" feature are created in a similar manner.

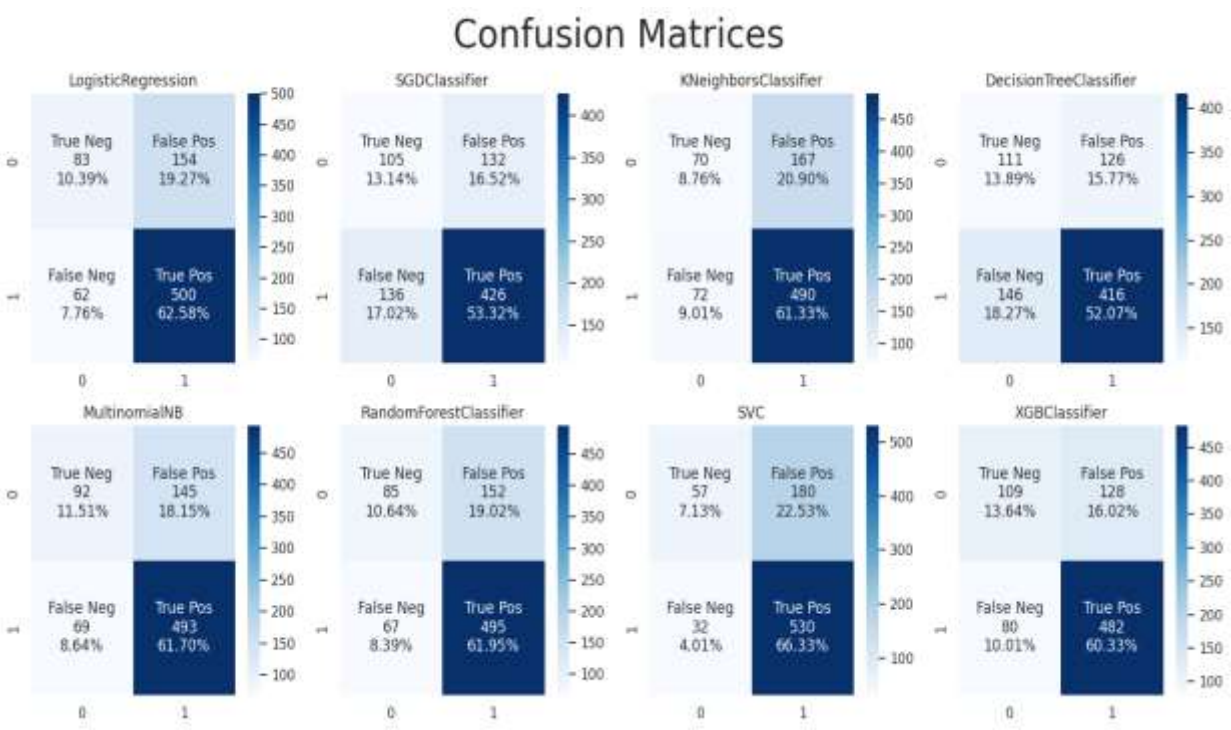
	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk	Age Category	Credit Category
0	67	male	skilled	own	none	little	1169	6	radio/TV	1	Senior	x < 2500
1	22	female	skilled	own	little	moderate	5951	48	radio/TV	0	Student	5000 <= x < 7500
2	49	male	unskilled and resident	own	little	none	2096	12	education	1	Adult	x < 2500
3	45	male	skilled	free	little	little	7882	42	furniture/equipment	1	Adult	7500 <= x < 10000
4	53	male	skilled	free	little	little	4870	24	car	0	Adult	2500 <= x < 5000

Let’s see what new information the dataset adds with these new features.



	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk	Age Category	Credit Category
95	58	male	skilled	rent	little	moderate	15945	54	business	0	Adult	15000 <= x < 17500
637	21	male	skilled	own	little	none	15653	60	radio/TV	1	Student	15000 <= x < 17500
818	43	male	highly skilled	own	little	little	15857	36	vacation/others	1	Adult	15000 <= x < 17500
887	23	male	skilled	own	little	moderate	15672	48	business	0	Student	15000 <= x < 17500
915	32	female	highly skilled	own	little	moderate	18424	48	vacation/others	0	Young	greater than 17500

A young and skilled individual is more likely to have a greater ability to repay a risky loan. However, considering the purpose of the loan, specifically for vacation, from a risk perspective, it is not advisable to approve loans exceeding 17,500 DM. Such loans would be considered high-risk. Therefore, it is reasonable to assume that the loan application should not be approved, and it is best to remove this particular instance from the dataset.



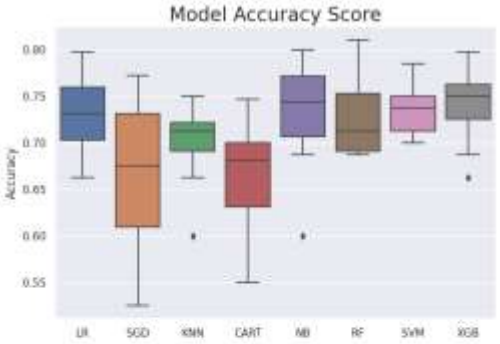
Accuracy is not the most suitable performance metric for classifiers, particularly when working with imbalanced datasets. A more effective approach to evaluating classifier performance is by examining a confusion matrix.

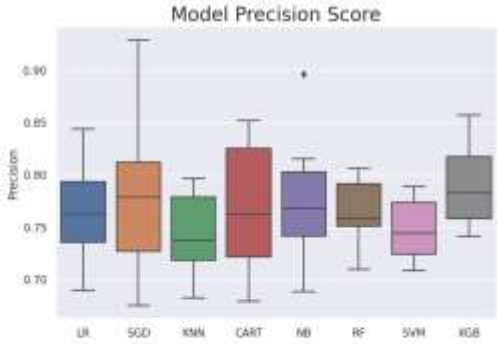
The concept behind a confusion matrix is to track the number of instances where class A is incorrectly classified as class B. In this matrix, each row corresponds to an actual class, while each column corresponds to a predicted class.

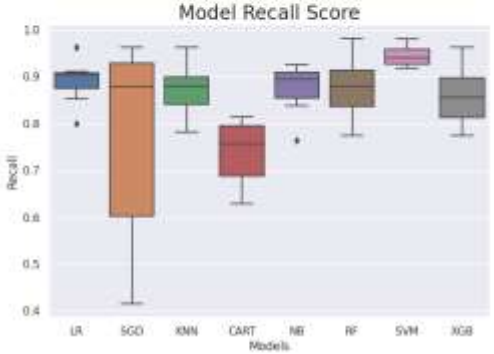
- Let's examine the initial confusion matrix, which serves as a performance measurement for the Logistic Regression model. In the first row of this matrix, we focus on bad loans (the negative class): 83 of them were accurately identified as bad loans (known as true negatives (TN)), while the remaining 154 were erroneously labeled as good loans (false positives (FP)).
- The second row pertains to good loans (the positive class): out of the total, 62 were incorrectly classified as bad loans (false negatives (FN)), while the remaining 500 were correctly identified as good loans (true positives (TP))."

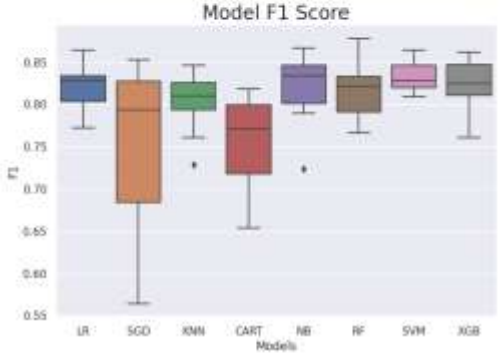
Models Comparison

Model	Accuracy	Precision	Recall	F1
LR	0.729747	0.765809	0.890437	0.822121
SGD	0.664715	0.783358	0.758134	0.744470
KNN	0.700934	0.745529	0.871866	0.802897
CART	0.659684	0.766212	0.738873	0.751476
NB	0.732231	0.773626	0.877772	0.821144
RF	0.726013	0.765198	0.880969	0.817776
SVM	0.734731	0.746576	0.943558	0.833106
XGB	0.739747	0.790736	0.857245	0.821147









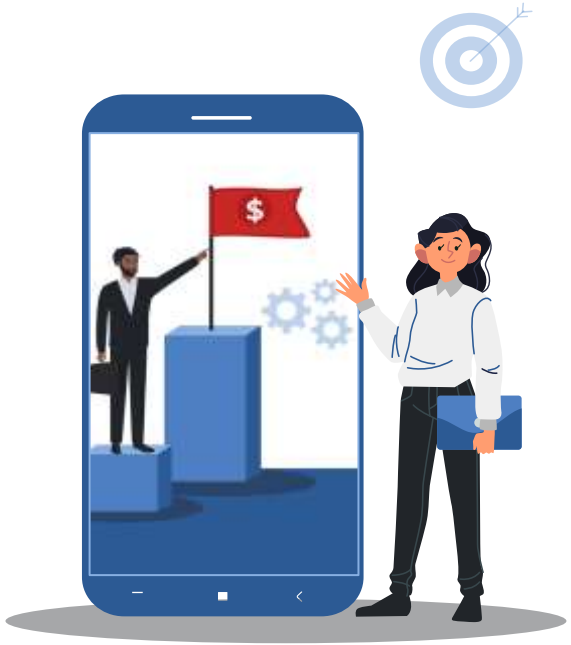
- Precision represents the proportion of accurately predicted positive observations compared to the total number of predicted observations. It is calculated using the formula:
- Precision = True Positives / (True Positives + False Positives)
- Recall, also known as sensitivity or true positive rate, measures the ratio of correctly predicted positive observations to the total number of actual positive class observations. The formula for recall is:
- Recall = True Positives / (True Positives + False Negatives)
- The F1 Score is a weighted average or harmonic mean of precision and recall. It is commonly used to compare two classifiers, as it favors classifiers that have similar precision and recall. Mathematically, the F1 Score is calculated as:
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

In our specific case study, we prioritize precision over recall. This is because approving a risky loan and subsequently incurring a loss of capital investment is more financially detrimental than missing out on the potential profits from a good loan.

Final model selection

Model	Accuracy	Precision	Recall	F1
XGB	0.71	0.743902	0.884058	0.807947

Model	Accuracy	Precision	Recall	F1
LogisticRegression	0.725	0.745562	0.913043	0.820847



Association of Categorical Features with Risk

Columns/ Features	Correlation/ Association	P-value
Sex (object)	5.348516218081436	0.020739913068713305
Job Category (object)	1.8851560280131707	0.5965815918843431
Housing (object)	18.19984158256362	0.00011167465374597684
Saving Accounts (object)	36.098928192418704	2.761214238568249e-07
Checking Account (object)	123.7209435162656	1.2189020722893755e-26
Purpose (object)	13.642086296939738	0.05792591119293625

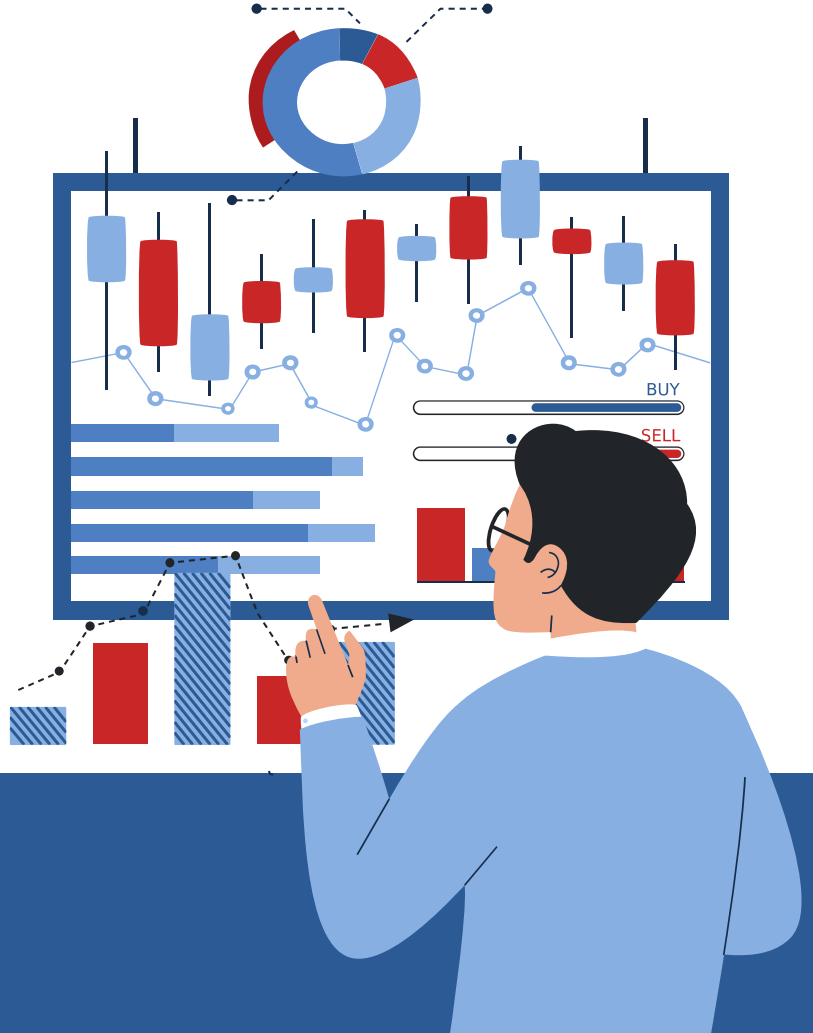
- Association and P-value of all the Categorical features collectively suggest that all the features are not directly related but they are statistically significant.
- There may be the possibility that collectively all the features are strongly associated to determine the risk of a customer being a defaulter.

Correlation of Numerical Features with Risk

Columns/ Features	P-value
Age (int)	0.0039253393982782 95
Credit Amount (int)	8.797572373533373e-07
Duration (int)	6.488049877187189e-12

- The p-value is very small (typically less than the conventional significance level of 0.05), it provides evidence against the null hypothesis, which is these columns cannot be used to determine the risk. Therefore, we can conclude that there is a statistically significant correlation between Risk and other numerical columns.
- There may be the possibility that collectively all the features are strongly associated to determine the risk of a customer being a defaulter.

Data Description



Description of the German credit dataset.

- Status of existing checking account
- Duration in month
- Credit history
- Purpose
- Credit amount
- Savings account
- Present employment since
- Installment rate in percentage of disposable income
- Personal status and sex
- Other debtors
- Present residence since
- Property
- Age in years
- Other installment plans
- Housing
- Number of existing credits at this bank
- Job
- Number of dependents
- Telephone
- Foreign worker

```
(1000, 21)
Class=1, Count=700, Percentage=70.000%
Class=2, Count=300, Percentage=30.000%
```


The dataset is loaded, and the number of rows and columns is verified to be 1,000 and 20, respectively. These 20 columns represent input variables, while there is one column dedicated to the target variable.Next, the class distribution is summarized, which involves confirming the count of good and bad customers, as well as the percentage of cases in both the minority and majority classes.

We can see that we have the correct number of rows loaded, and through the one-hot encoding of the categorical input variables, we have increased the number of input variables from 20 to 61. That suggests that the 13 categorical variables were encoded into a total of 54 columns.

Importantly, we can see that the class labels have the correct mapping to integers with 0 for the majority class and 1 for the minority class, customary for imbalanced binary classification dataset.

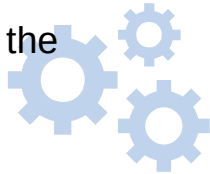
Next, the average of the F2-Measure scores is reported.

```
(1000, 61) (1000,) Counter({0: 700, 1: 300})
Mean F2: 0.682 (0.000)
```

In this case, we can see that the baseline algorithm achieves an F2-Measure of about 0.682. This score provides a lower limit on model skill; any model that achieves an average F2-Measure above about 0.682 has skill, whereas models that achieve a score below this value do not have skill on this dataset.

Precision is the ratio of correctly predicted positive observations to the total predicted observations

Recall (sensitivity or true positive rate) is the ratio of correctly predicted positive observations to the total observations of the actual positive class



Precision = TruePositives / (TruePositives + FalsePositives)

Recall = TruePositives / (TruePositives + FalseNegatives)

F1 Score is the weighted average or harmonic mean of precision and recall. F1 score favors classifiers that have similar precision and recall. It’s often used if you need a simple way to compare two classifiers.

F1 score = 2 * (precision * recall) / (precision + recall)

Now we are interested in the F-measure that will summarize a model’s ability to minimize misclassification errors for the positive class, but we want to favor models that are better are minimizing false negatives over false positives.

This can be achieved by using of the F-measure that which is same but favors higher recall scores over precision scores. This is called the Fbeta-measure, a generalization of F-measure, where “beta” is a parameter that defines the weighting of the two scores.

Fbeta-Measure = ((1 + beta^2) * Precision * Recall) / (beta^2 * Precision + Recall)

A beta value of 2 will weight more attention on recall than precision and is referred to as the F2-measure.

F2-Measure = ((1 + 2^2) * Precision * Recall) / (2^2 * Precision + Recall)

As we are iterating for many times, we are printing output as mean score and Standard deviation score of results :


```
>LR 0.498 (0.072)
```

In this case, we can see that the tested model have an F2-measure above the default of predicting the majority class in all cases (0.682). And our models are skillful. This is surprising, although suggests that perhaps the decision boundary between the two classes is noisy.

Undersampling is for imbalanced dataset refers to a dataset where the distribution of classes is highly skewed. Class imbalance occurs when the number of observations in one class (the minority class) is significantly lower than the number of observations in another class (the majority class). This can lead to biased models that perform poorly in predicting the minority class.


```
>ENN 0.706 (0.048)
>RENN 0.714 (0.041)
```

ENN is an undersampling technique that aims to remove instances from the majority class that are misclassified by the k-nearest neighbors classifier

RENN (Repeated Edited Nearest Neighbors):
RENN is an extension of ENN that applies the ENN algorithm multiple times iteratively. In each iteration, ENN is applied to the dataset, and the resulting dataset is used as input for the next iteration. This process continues for a specified number of iterations or until no further changes occur in the dataset.

After using undersampling techniques our improved and final result for logistic regression is :

```
0.743 (0.036)
```

Now we will apply our model on actual business problem, to make approval process better.

Applying model on our dataset,


```
Good Customers:
>Predicted=0 (expected 0)
>Predicted=0 (expected 0)
>Predicted=1 (expected 0)
Bad Customers:
>Predicted=1 (expected 1)
>Predicted=1 (expected 1)
>Predicted=1 (expected 1)
```

0 for “good customer”, and 1 for “bad customer”.

We are testing our model on some unseen data :

```
print('\n      :: LOAN APPROVAL ::\n')
unseen_data = [['A12', 36, 'A34', 'A41', 3200, 'A61', 'A72', 2, 'A92', 'A101', 3, 'A121', 55, 'A143', 'A152', 1, 'A173', 1, 'A192', 'A201'],
               ['A11', 24, 'A32', 'A41', 1500, 'A63', 'A71', 3, 'A94', 'A101', 1, 'A123', 25, 'A143', 'A152', 2, 'A172', 1, 'A191', 'A201'],
               ['A13', 48, 'A34', 'A41', 5500, 'A64', 'A71', 4, 'A93', 'A101', 2, 'A124', 30, 'A143', 'A152', 2, 'A173', 2, 'A191', 'A201']]
```

```
      :: LOAN APPROVAL ::

*Predicted=0
[Good customer, can give loan, can Offer Personal Loan to Customer]

*Predicted=1
[Bad customer, can't give loan & Customer Not Eligible for Personal Loan]

*Predicted=0
[Good customer, can give loan, can Offer Personal Loan to Customer]


       :: THANK YOU ::
```

THANK YOU!