
# CREDIT RISK ESTIMATION WITH MACHINE LEARNING AND ARTIFICIAL NEURAL NETWORKS ALGORITHMS

**Capstone Project**

**İlker Yıldız**

**İSTANBUL, 2021**

**MEF UNIVERSITY**

# CREDIT RISK ESTIMATION WITH MACHINE LEARNING AND ARTIFICIAL NEURAL NETWORKS ALGORITHMS

**Capstone Project**

**İlker Yıldız**

**Advisor: Asst. Prof. Dr. Berk Gökberk**

**İSTANBUL, 2021**

# MEF UNIVERSITY

Name of the project: Credit Risk Estimation with Machine Learning and Artificial Neural Networks Algorithms
Name/Last Name of the Student: İlker Yıldız
Date of Thesis Defense: 26/01/2021

I hereby state that the graduation project prepared by İlker Yıldız has been completed under my supervision. I accept this work as a "Graduation Project".

26/01/2021
Asst. Prof. Dr. Berk Gökberk

I hereby state that I have examined this graduation project by İlker Yıldız which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

26/01/2021

Director
of
Information Technologies Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

**THE EXAMINATION COMMITTEE**

| Committee Member | Signature |
|---|---|
| 1. Asst. Prof. Dr. Berk Gökberk | …….. |
| 2. …….. | …….. |

## Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

| Name | Date | Signature |
| --- | --- | --- |
| İlker Yıldız | 26.01.2021 | |

# EXECUTIVE SUMMARY

CREDIT RISK ESTIMATION WITH MACHINE LEARNING AND ARTIFICIAL NEURAL NETWORKS ALGORITHMS

İlker Yıldız

Advisor: Asst. Prof. Dr. Berk Gökberk

JANUARY, 2021, 28 pages

Credit risk assessment is very important for financial institutions today. The probability that a financial institution customer will not be able to repay the credits used is called credit risk. Financial institutions accept or reject credit applications. Institutions evaluate credit applications according to the personal information of the customers, life situation, loyalty, etc. If these data are below various values, financial institutions reject the application. The organization rejected the application because the client anticipated financial difficulties in the future. In the project, "German Credit" data on the Kaggle platform was used. In this data set, customers information and credit status are found as "good" and "bad". By using these data, it is aimed to evaluate new credit application requests. The data set used was passed through various pre-data processing steps and models such as Logistic Regression, Artificial Neural Networks, K-NN, Support Vector, Naïve Bayes, Decision Trees, Random Forest, LGBM and XGB were trained. The highest accuracy is achieved using the XGB model. (0.74)

**Key Words**: Credit Risk, Risk Analysis, German Credit Data, Machine Learning

# ÖZET

MAKİNE ÖĞRENMESİ VE YAPAY SİNİR AĞLARI ALGORİTMALARI İLE KREDİ RİSK TAHMİNİN YAPILMASI

İlker Yıldız

Proje Danışmanı: Dr. Öğr. Üyesi Berk Gökberk

OCAK, 2021, 28 sayfa

Günümüzde kredi risk değerlendirilmesi finans kuruluşları için çok önemlidir. Finans kuruluşu müşterisinin kullandığı krediyi geri ödeyememe olasılığına kredi riski denilmektedir. Finans kuruluşları kredi başvurularını kabul ya da red eder. Kurumlar kredi başvuruları müşterilerin kişisel bilgilerine, yaşam durumuna, kuruma olan sadakati vb. özelliklerine göre değerlendirir. Bu verilerin çeşitli değerlerin altında kalması durumunda finans kuruluşları başvuruyu reddetmektedir. Kuruluşun başvuruyu reddetmesinin nedeni müşterinin ilerleyen zamanda finansal olarak zorluk çekeceğini ön görmesidir. Projede kaggle platformu üzerinde bulunan "German Credit" verisi kullanılmıştır. Bu veri setinde müşterilerin bilgileri ve credit durumları "iyi" ve "kötü" olarak bulunmaktadır. Bu veriler kullanılarak yeni gelen kredi başvuru taleplerinin değerlendirilmesi hedeflenmektedir. Kullanılan veriseti çeşitli ön veri işleme adımlarından geçirilerek Lojistik Regresyon, Yapay Sinir Ağları, K-NN, Destek Vektör, Naïve Bayes, Karar Ağaçları, Rastsal Orman, LGBM ve XGB gibi modeleri eğitilmiştir. En yüksek doğruluğa XGB modeli kullanılarak erişilmiştir. (0.74)

**Anahtar Kelimeler**: Kredi Risk, Risk Analizi, Almanya Kredi Verisi, Makine Öğrenmesi

# TABLE OF CONTENTS

# LIST OF FIGURES

# 1. INTRODUCTION

## 1.1. About the Project

Credit risk assessment is a very important issue in financial institutions today. The probability that a customer will not be able to pay the credits received is called credit risk. A decision to granting or rejecting a credit is very important. Banks evaluate many features while giving credit to their customers. These features generally include customers credit history, life situation, loyalty and personal information. As per the statistical data obtained from federal reserve **[1]**, delinquency and charged off rates of banks of 2019 Q1 is 1.74 (Real estate loans) and 2.33(consumer loan category). This alarming situation in banks and financial institutions have sought the attention of various researchers. Modern data mining and machine learning techniques are widely used for credit risk estimation.

The German Credit data set was used in the project. The data set consists of 10 categories and 1000 observation units. Categories of data set; age, gender, job, housing status, checking account, credit amount, duration and risk. Customers are classified as "good" or "bad" credit risks according to these data. Customers classified as "good" repay their credit. Customers classified as "bad" could not repay the credit. Regarding the data, processes such as exploratory data analysis (EDA), missing value analysis, outlier analysis, rare analysis, feature engineering, encoding, feature scaling were applied. The data were processed after being made available for the models. Artificial Neural Network, Logistic Regression, K-NN, Support Vector Machine, Gaussian Naive Bayes, Decision Tree Classifier, Random Forest Classifier, L-GBM and XGB models were used in this project. The highest accuracy rate was achieved with the XBM model. (0.744)

## 1.2. Literature

Credit risk classification models are widely used today. There are classification sample data sets for credit risk. The most familiar of these datasets is German Credit Dataset. Various algorithms are used to classify customers in the data set according to credit risk. J. Kruppa et al. **[2]** some of these algorithms are Logistic Regression, K-NN, Support Vector Machine (SVM), Naive Bayes, Decision Tree, Random Forest Classifier, Gradient Boosting. Shi.L. et al. **[3]** Random Forest, SVM, Logistic Regression and K-NN algorithms were calculated according to accuracy and root mean square error. Confusion matrices were created as a result of the calculations. Random Forest algorithm has made the best credit risk conclusion. Malekipirbazari M et al. **[4]** as a result of the researches, it has been proven that Random Forest performs better than Logistic Regression and KNN. N. Arora et al. **[5]** Random Forest is ensemble classification algorithm which besides from providing good classification accuracy, also provides numerous other advantages such as immunity to over fitting, faster, simpler, helps in better estimating internal error and provides a good tolerance of outliers and noises. Bolasso is integrated with Random Forest, Naive Bayes, SVM, KNN and the best results are obtained with Random Forest. Y.Liu et al. **[6]** Generally, data sets on credit scoring are multidimensional. Therefore, the classification problem becomes more complex. As the size of the data set increases, the results of the prediction algorithms decrease. Each classification algorithm uses different feature selection methods according to the data. Therefore, its algorithms have their own advantages and disadvantages. As a result, different optimization trials increase the classification accuracy. Behr, A., & Weinblat, J. **[7]** The best classification result was obtained by Random Forest algorithm. In the literature search, it is seen that most researchers achieved high results with the Random Forest algorithm. As in other algorithms, not having over fitting problems increases the success problem. In this study, credit classification will be made using algorithms such as Logistic Regression, KNN, Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, ANN, XGBoost and LGBM. Considering other researches, success score will be tried to be exceeded.

# 2. ABOUT THE DATA

One of the most important services of banks that attract their customers is credit. If the prices of customers special needs has high prices, customers usually pay by credit. For this reason, many credit applications are made to banks. Banks compare various information of customers while giving credit. As a result of this comparison, it gives credit to the customer or not. Banks can give credit to customers in the long term. If the total repayment amount of the credit given in the long term is too high with the interest, sometimes customers cannot pay back. This situation causes the bank to money loss. Therefore, banks must sure that customers can repay their credit.

## 2.1. Features

This dataset is available on the kaggle platform. The aim is to determine the risk of a bank credit given to customers. The risk status of the customers for the new credit will be determined by using some machine learning, artificial neural network model with this dataset. The dataset includes the following features:

- ***Age:*** This variable bank contains the age information of the customers.
- ***Sex:*** This variable contains the gender information of the bank customers.
- ***Job:*** This variable contains occupational information of bank customers. Professions are divided into four categories according to skill level. (0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
- ***Housing:*** This variable includes the housing status of the bank customers.
- ***Saving Accounts:*** This variable bank contains customers savings information. Customers savings are classified as little, moderate, quite rich, rich.
- ***Checking Accounts:*** This variable show the cash in the accounts of bank customers.
- ***Credit Amount:*** This variable shows how much credit the bank customers get. DM - Dustsch Mark is used as currency.
- ***Duration:*** This variable contains the specified time to pay the credit.
- ***Purpose:*** This variable shows the purpose for which the bank customer taking credit. Taking credit by customers are classified as car, furniture / equipment, radio / TV, domestic appliances, repairs, education, business, vacation / others.

- ***Risk:*** This variable contains the customers risk information. Customers in the Data Set have previously taken credit from banks. A value of 1 was given for customers who paid the credit and 0 for customers who could not.

### 2.2. Exploratory Data Analysis (EDA)

In this section, the data set will be examined with visuals. The data set consists of many variables. Our dependent variable is "Risk" in the data set. Using other independent variables, the dependent variable is tried to be predicted. When the data set was examined, the distribution of the dependent variable was successful at a rate of seventy percent. In other words, the credit given to the customers by the bank has been repaid. However, thirty percent of customers did not make a refund. The data are shown in Figure 1.

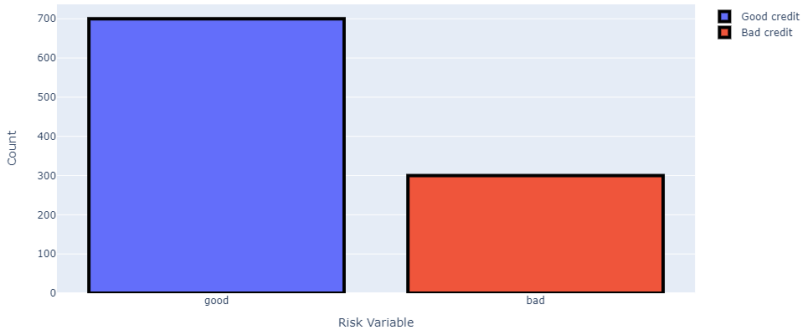


**Figure 1:** Target Value Distribution

There are many variables in the data set. One of the most important of these variables is the "Age" variable. Banks can legally give credit to their customers after reaching a certain age. Giving credit over a certain age is dangerous for banks. Health problems occur when people reach a certain age. This situation is reflected in their financial situation. In cases such as death, banks may not be able to receive a refund. For this reason, banks are afraid to give credit to too old customers. Young customers are usually students or new employee. The income of these customers is generally not fixed in a long term. Therefore, it is a risky group for banks. A frequency table was created for the age variable in the data set. According to this image, it was determined that customers between the ages of twenty and fourty have a high concentration of credit. It was found that the number of credits withdrawn decreases as the age increases. The data are shown in Figure 2.

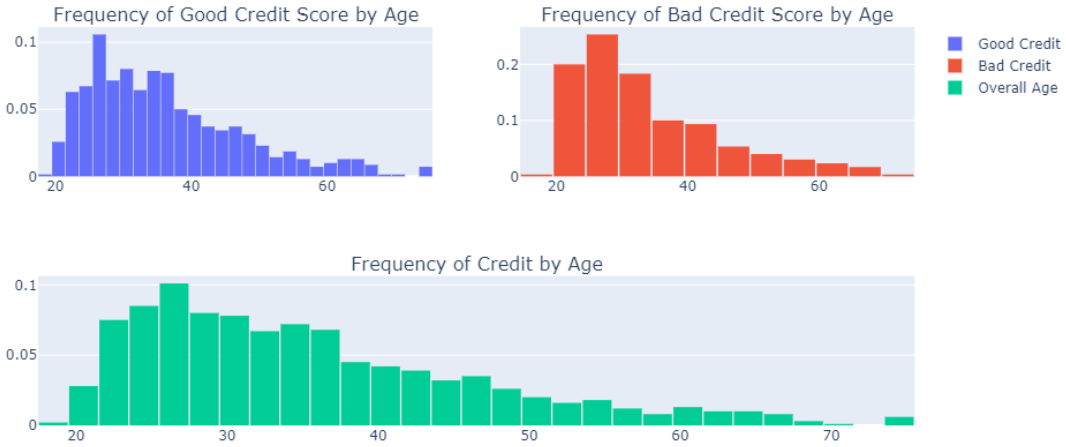


**Figure 2:** Age Frequency Tables

Another important variable is gender. When the gender variable is examined, there are 690 male and 310 female customers. According to these data, it can be said that male customers get more credits. 72% of male customers successfully paid credit and 64% of female customers successfully paid credit. Unsuccessful credits according to gender are generally in high majors. Male customers used a minimum of 276-DM and a maximum of 15945-DM credit. Female customers used a minimum of 250-DM and a maximum of 18424-DM credit. The data are shown in Figure 3.

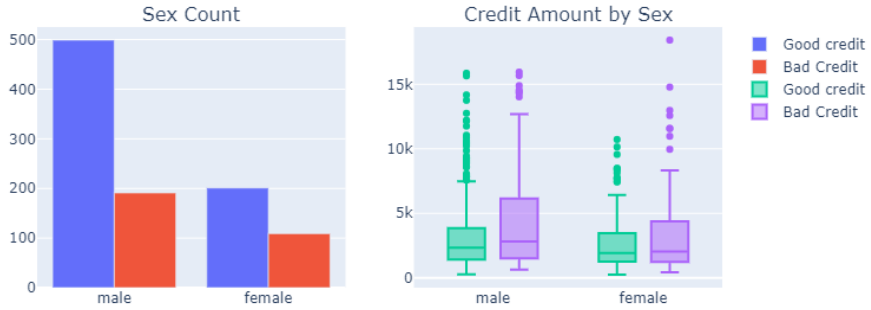


**Figure 3:** Sex Distribution

Another variable is job. Bank customers' income usually comes from jobs. Therefore, the business of bank customers is important. Customers pay their credits from this income. The higher the business class of the customers, the higher their income. Therefore, credit risks are decreasing. Professions are classified in 4 categories according to abilities in the data set. Level 0 skilled indicates unqualified customers, level 4 skilled indicates highly skilled customers. As the skill level of the customers increases, their income increases. According to the data set, the number of level 0 skilled customers is 22, the number of level 1 skilled customers is 200, the number of level 3 skilled customers is 630, and the number of level 4

customers is 148. It was observed that the 2nd level skilled customers received the most credit. Level 3 skilled customers attracted the highest amount of credits. The data are shown in Figure 4.

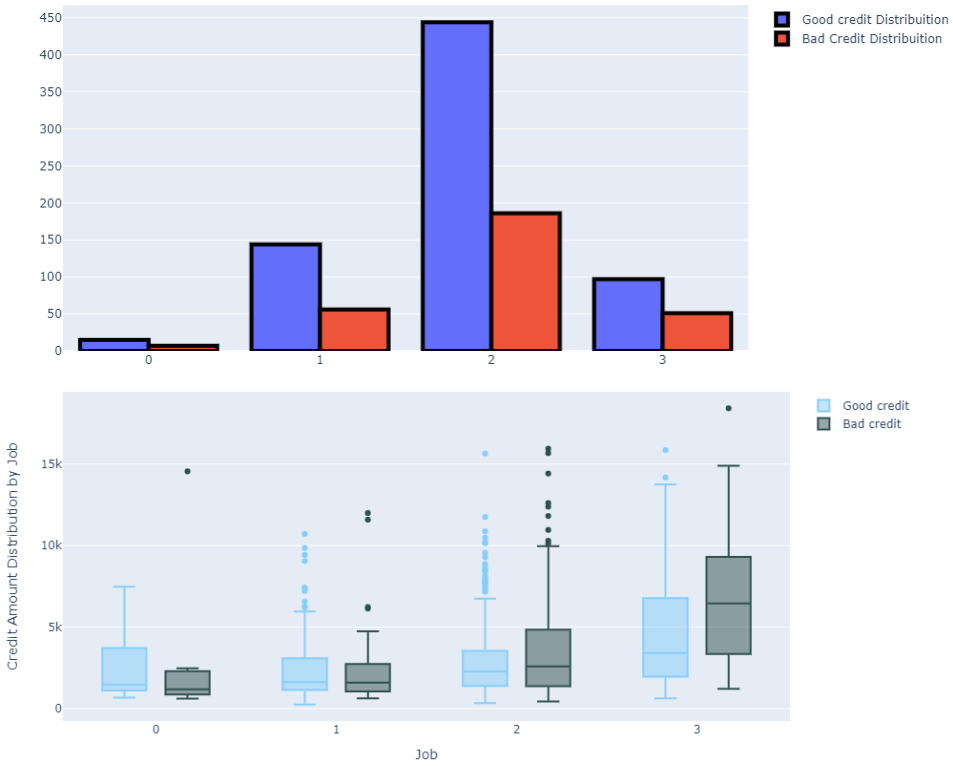


**Figure 4:** Job Distribution Frequency Table and Credit Amount Distribution by Job

Housing data shows the housing situation of the customers. Today, the biggest part of person expenses is housing. This is why the risk ratio of customers who own their home is low. They do not allocate a large part of their income for housing. In addition, if there is no payment, their homes can be mortgaged. Giving credit to customers who own a home is less risky. In the data set, it is observed that the rate of paying credit of home owners is higher than other groups. Customers are classified in 3 categories according to their housing status. These categories are own, rent and free. In the own category, customers live in their own home. In the rent category, customers live in rental houses. In the free category, customers have no homes. The number of customers who own a house is 713. This accounts for 71.3% of the entire data set. 17.9% of customers live in a rental house. Homeless customers make up 10.8% of the data set. 73% of the customers in the own status successfully repaid credit. 60% of customers with rent status successfully repaid credit. Customers in free status were successful by 59%. The data are shown in Figure 5.

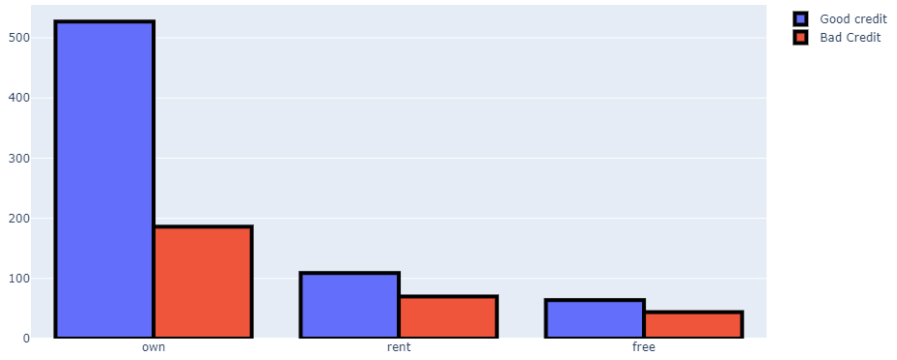


**Figure 5:** Housing Distribution

Another data category is saving account. This data shows the amount of money saved by bank customers. This data is analyzed in four sub-categories. These categories are little, moderate, quite rich and rich. The highest amount of credit is in the little category. The savings situation of the bank customers in the little category is low. Therefore, bank customers need credits when purchasing products. The number of customers in the little category is 603. This unit constitutes 60% of the total data set. 64% of the customers in the little category have paid their credits. The number of customers in the moderate category is 103. It constitutes 10% of the data set. 67% of the customers in the moderate category have paid the credits they used. The number of customers in the quite rich category is 63. This category constitutes 6% of the data set. 82% of customers in the quite rich category have refunded their credits. The number of customers in the rich category is 48. 87.5% of customers in the rich category have refunded their credits. When this category is examined, it has been determined that when the savings status of bank customers increases, the success rate increases. In addition, it has been determined that when the savings situation increases, the credit amount decreases. The data are shown in Figure 6.

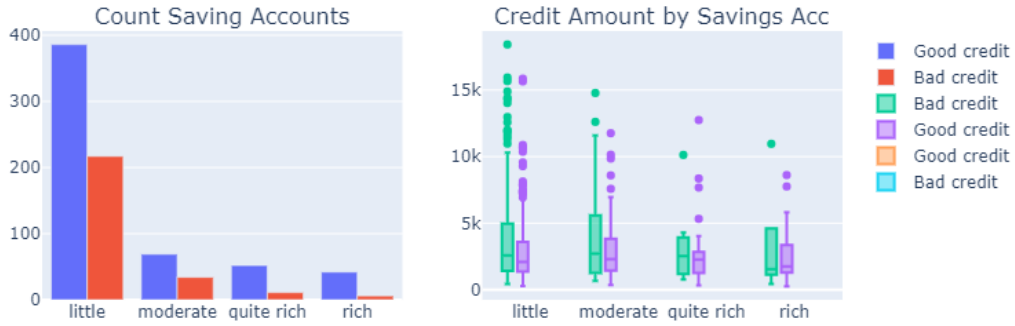


**Figure 6:** Count Saving Account and Credit Amount by Saving Account Graphs

Another variable in the data set is Credit Amount. This data shows how much credit customers get. It is seen that customers receive credit between 0 and 5000 DM at most. Figure 7 shows the frequency table for Credit Amount. The green unit represents bad credits. The yellow unit represents good credit. According to this chart, only credits between 0 and 5000 DM success rate is above 50%. The rate of payment success of the remaining credits is below 50%.

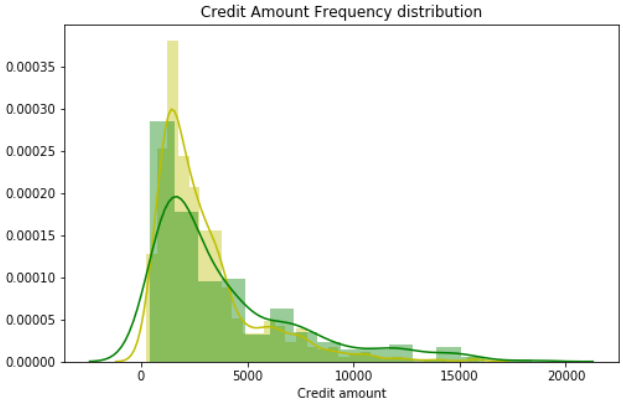


**Figure 7:** Credit Amount Frequency Chart

One of the variables in the data set is Duration. This data shows how many monthly credits customers receive. Customers typically received 6, 18, 24, 36 and 48 months of credit. The reason customers choose these months is that these months are multiples of twelve months. Looking at the data set, it was seen that the most credits were received in 12 and 24 months. It was successful in paying when customers received credit for at most 6 months. The reason for this is that the credit amount of the 6-month credits is low. Therefore, it can be said that customers pay the credits easily. It is seen in the data set that there is a problem in the payment of only 48-month credits. In all the remaining months, the customer success rate is over 50%. The data are shown in Figure 8.

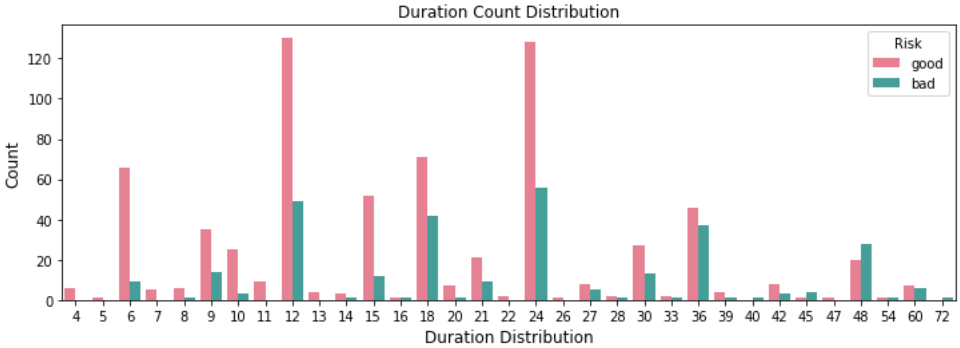


**Figure 8:** Duration Distribution Chart

The last variable in the data set is purposes. This data shows why customers use the credits. According to this data, customers mostly used credits for radio / TV, furniture or car. 28% of the bank customers received credit for radio / TV. 77.8% of customers who received credit for radio / TV successfully repaid the credit. 33.7% of the bank customers received credit for cars. 68.5% of customers who received credit for the vehicle successfully repaid the credit. 18.1% of the bank customers received credit for furniture. 67.9% of customers who received credit for furniture successfully repaid the credit. The bank used credit in the highest vacation / other category of its customers. Bank customers used credit in the lowest domestic appliances category. Most of the customers who use credit are between the ages of 20 and 40. All graphics related to Purposes are shown in Figure 9.

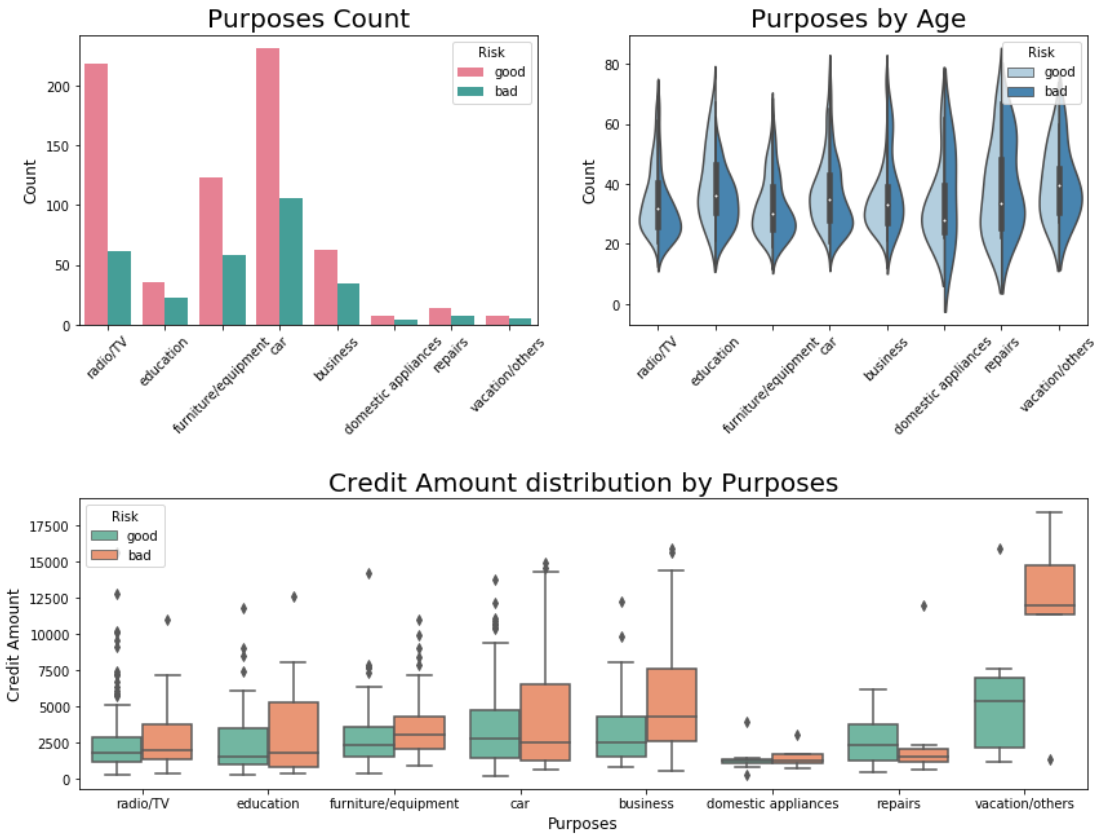


**Figure 9:** Charts for Purposes by Age, Purposes by Credit Amount and Purposes count

# 3. DATA PREPROCESSING

## 3.1. Missing Value Analysis

Missing value is the data that is missing from the data set due to various reasons. These data affect the accuracy rate in some machine learning models. Therefore, missing values should be eliminated. These values can be deleted or filled in. Before processing the data, it should be observed why there are missing values. Sometimes the value of the data can be left blank when it is zero. Therefore, attention should be paid to the description of the data source. Missing values can be filled in with values such as mode, median or mean.

When the data set is examined, the missing data are available in the saving account and Checking account categories. These categories contain numerical data. 183 missing data were found in the saving account category and 394 in the checking account category. These data were filled with mod according to gender, risk category and age values. The distribution of missing data is shown in the Figure 10.

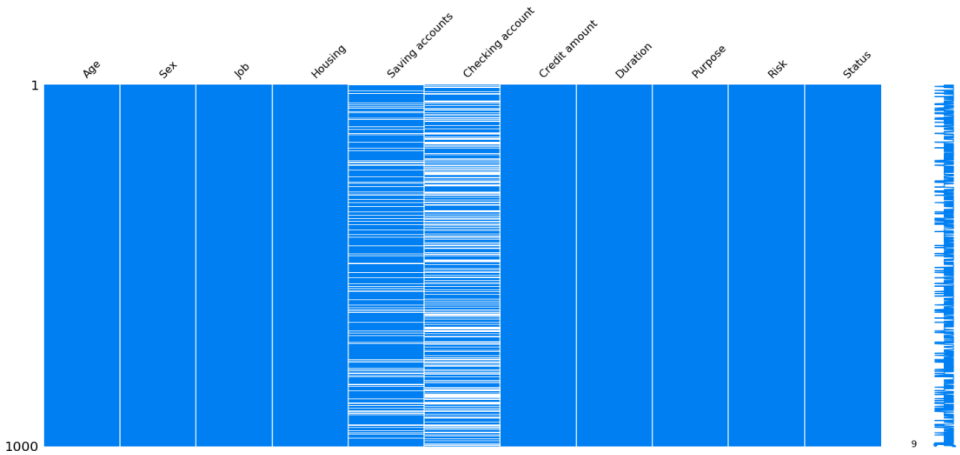


**Figure 10:** The distribution of missing data

## 3.2. Outliers Analysis

Outliers may arise due to errors such as measurement errors and registration errors. Outlier data can sometimes occur by adding very low probability events to the data. Outliers can drastically change the results of data analysis and statistical modeling. They increase the error variance and reduce the power of statistical tests. Outliers can lower normality if randomly distributed. For this reason, outlier data should be identified and processed from the data set with various methods. Boxplot method, z-score method or clustering method can be used for outlier detection.

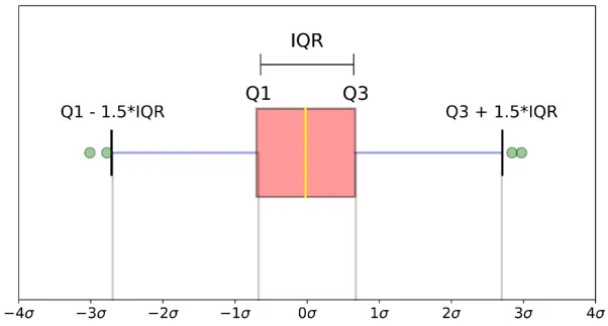


**Figure 11:** Box Plot

$$Q_1 = 1.5 * (Q_3 - Q_1)$$

$$Q_3 = 1.5 * (Q_3 + Q_1\ )$$

$$Inter\ Quantile\ Range\ (IQR) = Q_3 - Q_1$$

The boxplot method was used to find outlier observations in the data set. With this method, Outlier limits are determined to look for outliers. Data outside these limits are called outlier data. When our data set is examined, outlier observation analysis can be performed in Age, Credit Amount and Duration categories. Credit amount is the amount of credit customers have used. Therefore, it may contain high or very low data. Duration variable is the information of how many months customers will pay. Customers have several payment options available for credit term. The application of cross-observation analysis for the age variable is contradictory. There are many older customers. This situation can be very rare. Therefore, an outlier analysis should be applied to these variables. Observation analysis was applied against the specified categories. 5% and 95% values were chosen for quarter value ratios. It was concluded that there were no contradictory observations in the data according to the distribution.

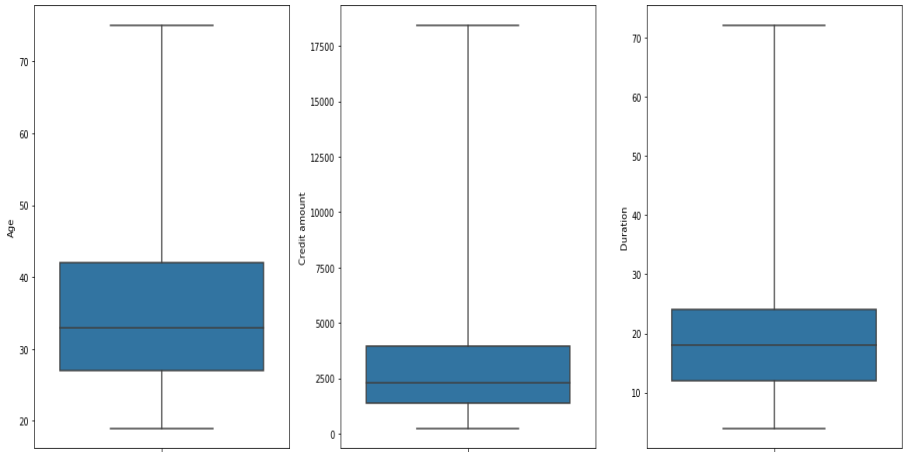


**Figure 12:** Box plot for Age, Credit Amount and Duration variables in outlier analysis

### 3.3. Rare Analysis

Rare analysis is called the processing of data that is rare in the distribution of data. Such data may be due to errors due to data entry. Therefore, the data source must be queried. If there is no problem due to data entry, rare data are events that are unlikely to occur. These data can be removed from the data set. However, if the removal process is implemented, valuable data will be deleted. In this case, the accuracy rate of the methods to be applied at the end of the process decreases. Therefore, combining these data with other categories does not cause data loss. When the data set is examined, there is rare data in the Duration and Purpose categories. Duration is the payback period of credit taken. Customers demand repayments in different months from each other. Usually the infrequently claimed months are 13, 14, 16, 20, 22, 26, 40, 45, 47, 54 and 72. These data usually appear 1 or 5 times in the data set. Therefore, these rare data will be eliminated by deriving a new variable with these variables. Purpose is why credit is taken. Domestic appliances categories appear 12 times in the entire data set. These data were also combined with other categories.

### 3.4. Feature Engineering

Feature Engineering is the process of extracting attributes from raw data. Machine learning algorithms are given observation units. After this data is processed, the algorithms make a label estimation. Therefore, it is important to extract unused data from raw data. First, rarely used categorical data should be removed from the data set. For this reason, rare analysis is done. As a result of the rare analysis, rare data were found in the Duration and Purpose categories. Duration variable indicates how long customers will pay their credits. This variable has been converted to year and named as integer. At the end of this transformation, rare data has been removed. One of the important variables in the data set is the age of the customers. When these data are examined, there are rare data. For this reason, this data group is also categorically classified. Finally, the process has been applied on the Credit amount data. Customers were categorically divided according to the credit they received. Data changes are shown in Figure 12. As a result of transforming these variables, the success rate increased by 3.7%.

| Duration | |
|---|---|
| Value | New Value |
| 01 – 12 | 0 – 1 Year |
| 13 – 24 | 1 – 2 Year |
| 25 – 36 | 2 – 3 Year |
| 36 – 48 | 3 – 4 Year |
| 49 – 60 | 4 – 5 Year |
| 61 – 72 | 5 – 6 Year |
| 73 – 84 | 6 – 7 Year |

| Age | |
|---|---|
| Value | New Value |
| 18 – 25 | Student |
| 26 – 35 | Young |
| 36 – 60 | Adult |
| 61 – 84 | Senior |

| Credit Amount | |
|---|---|
| Value | New Value |
| 250 – 4793 | Poor |
| 4794 – 9336 | Mid |
| 9337 – 13879 | Upper |
| 13880 – 18424 | Rich |

**Figure 13:** Representation of the new variables generated

### 3.5. Feature Scaling

The dispersion of data is a factor that affects the operation of some algorithms. Model performance can be affected if data is right or left tilted. For algorithms that use distance-based calculations such as Euclid, Manhattan, values will deviate. In this context, we can obtain more accurate results by pulling these features into a common data range. There are some methods to normalize these values and reduce dominance. These are methods such as normalization and standardization. These methods are useful methods to apply before using distance-based and gradient-based estimator algorithms.

***MinMax Scaling*** is a situation in which data takes values between 0 and 1. The distribution is similar to the distribution of the data. In this method, there is sensitivity against outlier data. For this reason, it may not perform well in cases where outlier data is excessive.

***Max-Abs Scaler*** scales and transforms each property individually, with a maximum absolute value of 1 for each property.

***Standardization*** is a method where the mean value is 0 and the standard deviation is 1, and the distribution approaches the normal. The data is subtracted from the mean value and then divided by the variance.

***Robust Scaler*** works similarly to normalization. It may give better results in data with outliers. It is similar to the data distribution. However, outliers are left out. The median value is removed for later use. The values are assigned to the 1st and 3rd quartile range. Scaling was applied for numeric variables using the robust scaler method in the data set.

# 4. MODELS

## 4.1. Logistic Regression

Machine learning models are often used to classify collected data or to predict new data. Prediction algorithms are used for numerical data and classification algorithms are used for categorical data. Logistic Regression is an algorithm for classification. It can easily classify categorical and numerical data. It works depending on the dependent variable. Therefore, the dependent variable must be encoded as binary (1 or 0). It is widely used in linear classification problems. Therefore, it is very similar to the Linear Regression model. The purpose of Logistic Regression is to define the relationship between the dependent variable and the independent variables in the data.

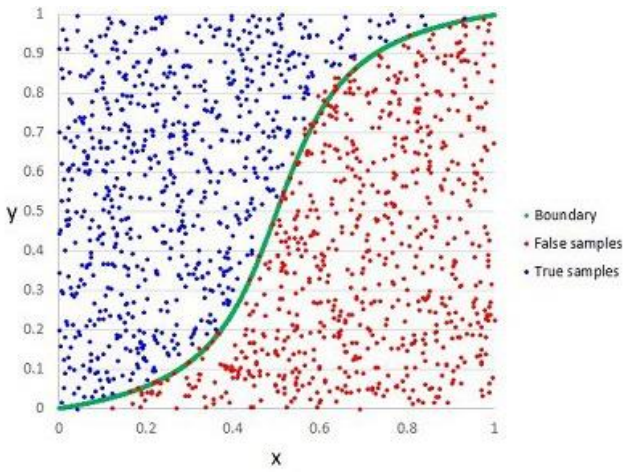


**Figure 14:** Visualization of Logistic Regression

There are situations that need to be considered when using Logistic Regression. Data that have nothing to do with the dependent variable should be removed from the model. This unnecessary data reduces the speed and accuracy of the model. Therefore, duplicate and incomplete data should be removed from the data set. The measurement rate of independent variables affects the accuracy of the model. Therefore, the error rate of the data set should be low. The model gives insufficient results in data sets with high error rate. The sample size should include sufficient data for model health. Despite these, Logistic Regression has many advantages. Parameters can be easily interpreted. Easy-to-use functions are produced mathematically.

### 4.2. K- Nearest Neighbor (KNN)

The K-NN algorithm is one of the simplest and most used algorithms. It is used in both classification and regression problems. K-NN algorithm is a lazy and non-parametric learning algorithm. It does not perform learning steps like many algorithms. Instead, it looks for the closest neighbors in the entire data set. It classifies as much as the value to be given to it. Analytical monitoring is possible. Resistant to noisy training data. Despite these advantages, it requires a large amount of memory space because it keeps the distance information for all points in memory. If the size of the data set increases, it significantly increases the cost and decreases the performance.

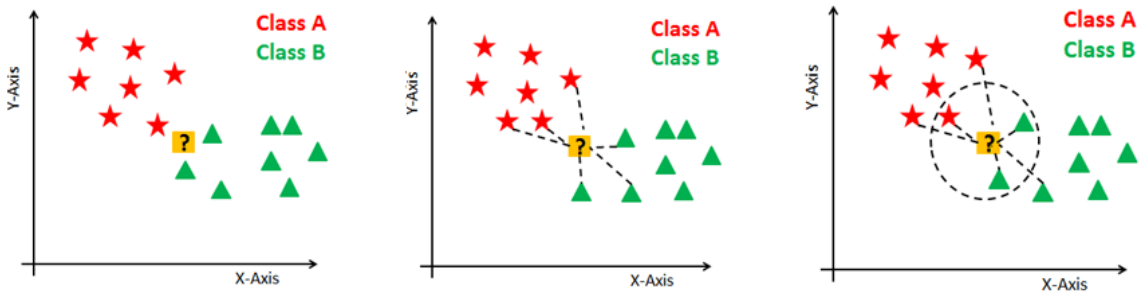


**Figure 15:** Visualization of K-NN

The K value is determined before running the algorithm. The meaning of this K value is the number of elements to look at. When a value comes, the distance between the value is calculated by taking the nearest K number of elements. The Euclidean function is generally used in the distance calculation. Manhattan, Minkowski and Hamming functions can also be used as an alternative to the Euclidean function. After the distance is calculated, it is ranked in ascending order and the corresponding value is assigned to the appropriate class.

### 4.3. Support Vector Machine (SVM)

Support Vector Machines (SVM) is a supervised learning algorithm based on the statistical learning theory. Support Vector Machines are mainly used to distinguish between two classes of data in the most convenient way. For this, decision boundaries or in other words hyper planes are determined. Support Vector Machines algorithm is used in many classification problems such as face recognition systems and voice analysis. Support Vector Machines are divided into two according to the linear separation and non-linear separation of the data set.

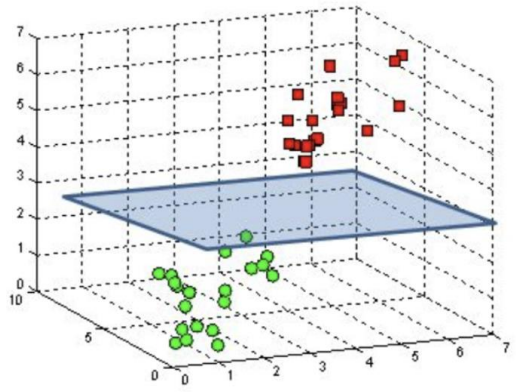

**Figure 16:** Visualization of Support Vector Machine

**Linear Support Vector Machines:** In classification with support vector machines, samples belonging to two classes are assumed to be linearly distributed. In this case, it is aimed to separate the two classes in the data set with the help of a decision function obtained using the training data. The line that divides the data set into two is called the decision line. Although it is possible to draw infinite decision lines, the important thing is to determine the optimal decision line. In order for the decision line to be resistant to the newly added data, the border line must be at the closest distance to the border lines of the two classes. The points closest to this border line are called support points. Class labels in the form of (-1, + 1) are generally used in classification with support vector machines.

**Nonlinear Support Vector Machines:** SVM Algorithm cannot draw a linear hyper plane in a nonlinear data set. For this reason, kernel tricks called kernel numbers are used. The kernel method highly increases machine learning in nonlinear data. The most commonly used kernel methods are Polynomial Kernel and Gaussian RBF (Radial Basis Function) Kernel.

### 4.4. Naive Bayes Algorithms

Bayes theorem is an important topic studied in probability theory. This theorem shows the relationship between conditional probabilities and marginal probabilities within the probability distribution for a random variable. The Naive Bayes classifier is based on Bayes' theorem. It is a lazy learning algorithm. The algorithm calculates the probability of each state for an element and classifies it according to the one with the highest probability value. Naive Bayes can achieve very successful training results with little training data. In some cases, the algorithm may not be able to observe a value in the test set in the training set. For this reason, the algorithm returns zero as a result of the probability value, so the algorithm cannot predict. This condition is commonly known as Zero Frequency. Correction techniques can be used to resolve this situation. One of the simplest correction techniques is known as Laplace estimation. There are many Naive Bayes implementations. These are;

***Gaussian Naive Bayes:*** If the data is continuous value, it is assumed that these values are sampled from a Gaussian distribution. In other words, it is assumed to be sampled from a normal distribution.

***Multi-nominal Naive Bayes:*** This method is used to classify data with more than one class.

***Bernoulli Naive Bayes:*** This method makes classification similar to Multi-nominal Naive Bayes. However, the prediction results are only boolean (binary)

The Naive Bayes method has many advantages in the field of machine learning. Each feature is considered independent of each other. Therefore, it performs better than models such as Logistic Regression. Simple and easy to apply. Good learning outcomes can be achieved with little data. The Method can be used with continuous and discrete data. It can be used with continuous and discrete data. It can also be used in an unbalanced data set. It can work well on high dimensional data. It can be used in real time systems due to its speed. The biggest drawback of the Naive Bayes method is that every feature is dependent on each other at some point in real life. Therefore, the model cannot model the relationships between variables.

Naive Bayes method; real-time systems are widely used in areas such as multiple classification problems (News / E-Commerce Categories), text classification (Spam Filtering / Sentiment Analysis), disease diagnosis and advice systems.

### 4.5. Decision Tree Algorithms

Tree-based learning algorithms are among the most used supervised learning algorithms. The algorithm can generally classify and regress all problems. Methods such as decision trees, random forest, gradient boosting are widely used in all kinds of data science problems. Therefore, it is very important for data analysts to learn and use these algorithms. Decision tree algorithm is one of the data mining classification algorithms. They have a predefined target variable. It offers a top-down strategy in terms of their structure. A decision tree is a structure used to divide a data set containing a large number of records into smaller sets by applying a set of decision rules. In other words, it is a structure that is used by applying simple decision-making steps and dividing large amounts of records into very small groups of records. Algorithm selection is based on the type of target variable. The most frequently used algorithms in decision trees; Entropy, Gini, Classification Error for categorical variables; for continuous variables it is the Least Decision method.

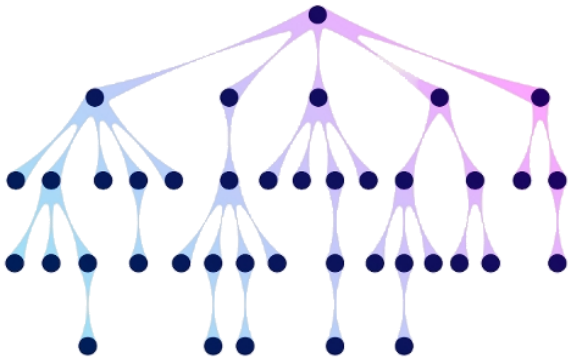

**Figure 17:** Decision Tree Visualization

The decision tree method has many advantages. The results of decision trees are easy to understand and interpret. The tree structures used can be visualized. Needs little data preparation. Therefore, it can be easily applied to small data sets. They are affected by missing data. Missing data must be filled in or removed as decision trees are affected by missing data. It can process both numerical and categorical data so they can handle multi output problems. A model can be validated using statistical tests. Decision tree algorithm does not have an approach to space distribution and classification structure.

### 4.6. Random Forest Algorithms

The Random Forest algorithm is a controlled classification algorithm. The algorithm simply creates a forest from decision trees. There is a direct proportion between the number of decision trees and the result in the algorithm. If the number of decision trees created increases, the result is getting closer to the truth. There are differences between the Random Forest algorithm and the Decision Tree algorithm. The main difference between the two algorithms is that the root node discovery and splitting operations work randomly. Random Forest adds additional randomness to the model while growing trees. Instead of looking for the most important feature when breaking down a node, it looks for the best feature among a random subset of features. Therefore, it results in a better score and a wide variety.

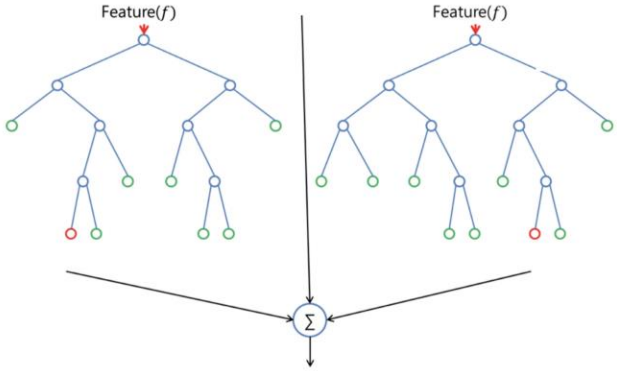


**Figure 18:** Random Forest Algorithm Visualization

The Random Forest Algorithm can be used for both classification and regression operations. There is an over fitting problem for all machine learning algorithms. If the number of trees created as a result of the Random Forest algorithm is sufficient, the over fitting problem does not affect the result. In addition, the rate of exposure to random forest algorithm missing data is very low. The Random Forest algorithm first takes the test properties. Then, the rules of the randomly generated decision tree are used to predict the results and store the predicted result. Scoring is calculated for each estimated result. Finally, the prediction with the highest vote is selected from the Random Forest algorithm. Random Forest algorithm is actively used in Banking, Medicine, Stock Exchange and E-commerce systems.

### 4.7. Boosting for Decision Tree Algorithms

The purpose of the boosting algorithms is to make inferences from the collection of trees obtained by giving different weights to the data set. Initially, all observations have equal weight. As the tree community begins to grow, the weights are rearranged according to the problem situation. While the weight of incorrectly classified observations is increased, the weight of observations that are rarely misclassified is reduced. In this way, trees gain the ability to organize themselves in difficult situations. Some of the boosting algorithms are GBM, XGBoost and Light GBM.

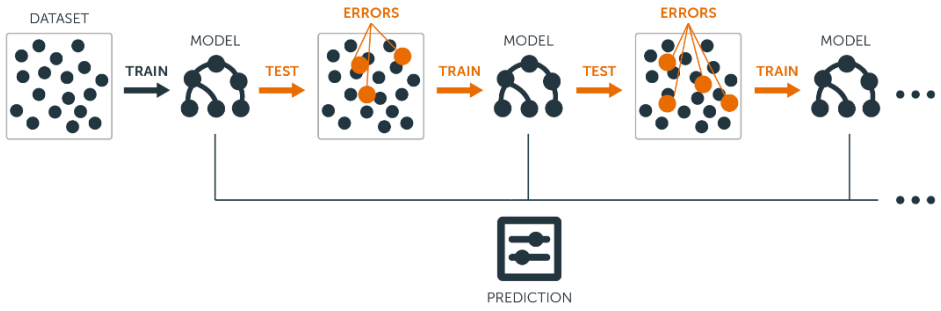


**Figure 19:** Boosting Algorithm

#### 4.7.1. Gradient Boosting (GBM)

Gradient Boosting or GBM is a machine learning algorithm that works for both regression and classification problems. GBM uses an escalation technique, combining a number of weak learners to form a strong learner. Regression trees used as basic learners, the generated tree is based on errors computed by the previous tree. With the rapid increase in data size and variety in recent years, the importance given to algorithm optimizations is increasing.

For this reason, as an alternative to the Gradient Boosting algorithm, algorithms that can accept versions of Gradient Boosting such as XGBoost, LightGBM, Catboost have been developed. It is aimed to achieve faster training and higher accuracy with these algorithms.

### 4.7.2. XGBoost

Tianqi Chen and Carlos Guestrin have been involved in our lives with the article "XGBoost: A Scalable Tree Boosting System" published in 2016. The most important features of the algorithm are that it can obtain high predictive power, prevent over-learning, manage empty data and do them quickly. Software and hardware optimization techniques have been applied to achieve superior results using less resources. Shown as the best of decision tree based algorithms. The working logic is very similar to Gradient Boosting.

The first step in XGBoost is to make the first prediction (base score). This estimate can be any number as the correct result will be reached by converging with the operations to be done in the next steps. How good this prediction is examined with the residual estimates of the model. Errors are found by subtracting the estimated value from the observed value.

After determining which tree has the higher earning value and deciding to use the tree, the prune process will begin. The value called "gamma" is chosen for pruning. Gamma is an assessment brought to the earning score. Branches with a gain score lower than the gamma score will be pruned. Therefore, increasing the gamma only helps to keep valuable branches in the tree and prevent over learning. Pruning continues from the last branch upwards. If it is decided not to prune the lowest branch, there is no need to investigate the branches above it. Afterwards, the trees are calculated as in the GBM model and the predictions continue to be corrected. These processes will continue until the errors are very low or the specified number of trees is reached. There are many parameters for XGBoost model settings. Some parameters used in the model; n_estimators are subsample and max_depth. n_estimators; the number of trees to be installed in the model, subsample; The line rate taken to create each tree, max_depth represents the depth of the tree. Higher success rates can be achieved by optimizing the parameters used in the model.

### 4.7.3. Light GBM (LGBM)

LightGBM is a boosting algorithm developed in 2017 as part of the Microsoft DMTK (Distributed Machine Learning Toolkit) project. Compared to other boosting algorithms, it has advantages such as high processing speed, ability to process large data, less resource (RAM) usage, high prediction rate, parallel learning and support for GPU learning. LightGBM is a histogram-based algorithm. The algorithm makes continuous data discrete. Therefore, it reduces the cost. The training time of decision trees is directly proportional to the number of calculations and divisions made. This method reduces training time and reduces resource usage. Two strategies can be used in learning decision trees. These are divided into level oriented and leaf oriented. In the level focused strategy, the balance of the tree is maintained while the tree structure grows.

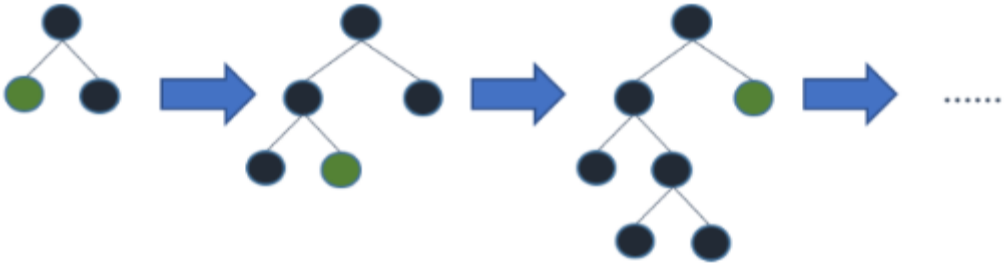

**Figure 20:** Leaf focused strategy tree growth

In the leaf-focused strategy, the division process from the leaves continues, which reduces the loss. For this reason, Light GBM algorithm differs from other boosting algorithms. The model has less error rate and learns faster with leaf-focused strategy. However, the leaf-focused growth strategy causes the model to be prone to over-learning when the number of data is low. Therefore, the algorithm is more suitable for use in big data. In addition, if parameters such as tree depth and leaf number are optimized, over-learning can be prevented.

**Figure 21:** Level-oriented strategy tree growth

LightGBM also uses two techniques different from other algorithms. These techniques are Gradient Based One Way Sampling and Special Variable Package. In addition, these techniques make calculations regarding the number of data samples and variables.

**Gradient-based One-Side Sampling:** GOSS aims to reduce the number of data without changing the accuracy of decision trees. Traditional Gradient Boosting scans all data samples to calculate information gain for each variable. However, GOSS only uses important data. Thus, the number of data is reduced without much affecting the distribution of the data.

**Exclusive Feature Bundling:** EFB aims to reduce the number of variables without changing to accuracy. Accordingly, it aims to increase the efficiency of model training. EFB has two process steps. These are creating packages and combining variables in the same package. EFB sparse features are combined to create more intense features. Accordingly, it leads to a decrease in complexity and faster training with lower memory consumption.

In summary, GOSS reduces data size to compute knowledge gain by neglecting less important data. EFB combines variables to reduce dimensionality. With these two functions, LightGBM increases the efficiency of the training process. Learning_rate, max_dept, num_leaves, min_data_in_leaf parameters can be optimized to prevent excessive learning in LightGBM. Feature_fraction, bagging_fraction and num_iteration parameters can be optimized to speed up learning time. Higher success rates can be achieved by optimizing the parameters used in the model.

### 4.7.4. Artificial Neural Network (ANN)

Artificial neural networks use the learning methods of the human brain. Artificial neural networks consist of many cells. These cells work simultaneously to solve complex problems. Artificial neural networks have learning abilities. Artificial neural networks can recognize and classify patterns, and complete missing patterns. They can work in parallel and process real-time information. Artificial neural networks are mainly used in areas such as diagnosis, classification, prediction, control, data association, data filtering and interpretation. They have fault tolerance. They can work with incomplete or ambiguous information. In

faulty cases, they show graceful degradation. An artificial nerve cell consists of five parts; These parts are inputs, weights, addition function, activation fuction and outputs.

***Inputs:*** Inputs are data coming to neurons. Data from the inputs are sent to the neuron nucleus to be collected, as in biological nerve cells.

***Weights:*** The information coming to the artificial nerve cell is multiplied by the weight of the connections before reaching the nucleus through the inputs and transmitted to the nucleus. In this way, the effect of the inputs on the output to be produced can be adjusted.

***Addition Function (Merge Function):*** Data comes to the artificial neural cell by multiplying by their weight. The addition function collects this data and calculates the cell's net input.

***Activation function:*** This function takes the weighted sum of all inputs in the previous layer. It then generates an output value. (for example, ReLU or sigmoid).

***Outputs:*** The value that comes out of the activation function is the output value of the cell. Although each cell has more than one input, it only has one output. This output can be linked to any number of cells.

Artificial neural networks are structures formed by connecting artificial nerve cells to each other. Artificial neural networks are examined in three main layers; input layer, hidden layers and output layer. Information is transmitted to the network through the input layer. Afterwards, the information is processed in intermediate layers and sent to the output layer. Processing information; It is the conversion of information coming to the artificial neural network into output by using weight values. The network must be able to produce the correct outputs for the inputs. Therefore, the weights must be correct values. A multilayer neural network is shown in Figure 22.

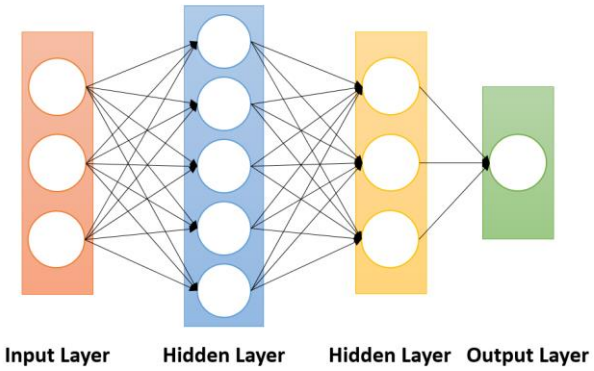


**Figure 22:** Multilayer Neural Network

# 5. RESULTS

The data set is taken over the kaggle platform. The data set was previously removed by the person sharing unnecessary categories. For this reason, feature selection has not been performed. The data set consists of 10 independent and 1 dependent variable. The arguments are age, sex, job, housing, saving accounts, checking account, credit amount, duration, purpose. The dependent variable is the risk category. There are 1000 different observation units in the data set. Figure 23.

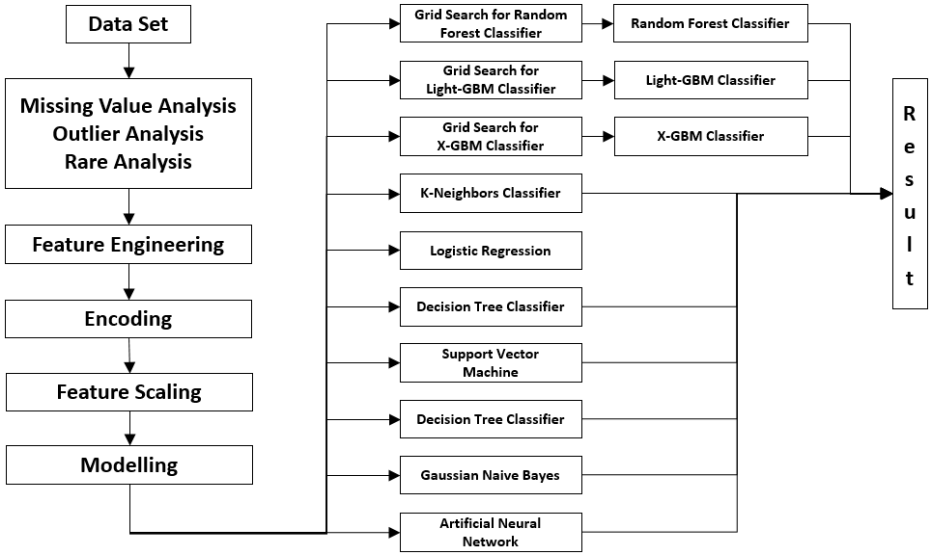


**Figure 23:** The graph of the transactions made is shown.

Exploratory data analysis was applied to the data set first. Information was tried to be obtained about this data set. The data has been made more understandable to visuals. As a result of this study, the following information was learned. According to the data, 70% of customers using credit from banks made a refund. 30% of the bank customers have not made a refund. The age range with the most credits is 20 to 40. As the age of the customers increases, the number of credits given by the banks decreases. 69% of customers using credits are men and 31% of customers are women. Bank customers are classified according to their housing status. It was observed that customers with houses received more credit and the payment rate was higher. Customers who get the most credits from customers have difficulty in saving money. According to the data set, the most used credit amount is between 250-5000 DM. The preferred period for paying credits is usually 6 - 12 - 15 - 18 - 24 - 36 and 48 months. Bank customers often used credit for radio / TV or cars.

Missing value analysis was applied to the data set. As a result of this analysis, 394 missing data were found for the saving account category and 183 for the checking account category. These missing data were replenished using gender, risk and age categories. Outlier observation analysis was applied to the data set. Quartile values of 0.05 and 0.95 were used to capture outliers. Outlier data could not be found because there is no data other than these quarter values. After these analyzes, feature engineering was applied and 3 new variables were derived. The generated variables were added to the data set and converted with the encoder. The data were made ready for models using the Robust scaler method.

Random forest classifier, Light-GBM classifier, xgboost classifier, kNN, logistic regression, decision tree classifier, support vector machine, gaussian naive bayes, artificial neural network models were used for data analysis. Parameter values of random forest, light-gbm and xgboost models were optimized. The gridsearchcv method was used for this. Gridsearchcv creates a separate model for each parameter to be tested in the model. The most successful hyperparameter set is determined according to the specified metric. A 3-layer artificial neural network model was created to process the data set. ReLu was used for the activation function of the first two layers. A sigmoid enable function was used for the last layer.

The bank shared information about its customers past credit transactions. The aim of the project is to evaluate new credit applications using this information. For this reason, the data was processed with 8 different machine learning algorithms and 1 artificial neural network models. The highest success rate was obtained with the XGBoost model. (0.744) The results of all models are shown in Figure 24.

| Models | | Result |
|---|---|---|
| Random Forest Classifier | RF | 0.712 |
| Light-GBM Classifier | LGBM | 0.716 |
| X-GBM Classifier | XGB | 0.744 |
| K-Neighbors Classifier | KNN | 0.664 |
| Logistic Regression | LGR | 0.696 |
| Decision Tree Classifier | DTC | 0.672 |
| Support Vector Machine | SVM | 0.652 |
| Gaussian Naive Bayes | GNB | 0.644 |
| Artificial Neural Network | ANN | 0.705 |

**Figure 24:** Results

## REFERENCES


**[1-5]** Nisha Arora, Pankaj Deep Kaur, 2019, A Bolasso based consistent feature selection enabled random Forest classification algorithm: An application to credit risk assessment, Applied Soft Computing Journal

**[3]** L. Shi, Y. Liu, X. Ma, Credit assessment with random forests, in: Emerging Research in Artificial Intelligence and Computational Intelligence, 2011, pp. 24–28

**[4]** M. Malekipirbazari, V. Aksakalli, Risk assessment in social lending via random forests, Expert Syst. Appl. 42 (2015) 4621–4631, Sciencedirect.

**[6]** X. Huang, X. Liu, Y. Ren, Enterprise credit risk evaluation based on neural network algorithm, Cogn. Syst. Res. (2018)

**[7]** A. Behr, J. Weinblat, Default patterns in seven EU countries: A random forest approach, Int. J. Econ. Bus. 24 (2) (2016) 181–222

**[8]** P. Danenas, G. Garsva, Selection of support vector machines based classifiers for credit risk domain, Expert Syst. Appl. 42 (6) (2015) 3194–3204, http://dx.doi.org/10.1016/j.eswa.2014.12.001.

**[9]** Hatipoğlu, E. (218, 12 July). *Machine Learning — Classification — Logistic Regression — Part 8*. Medium: https://medium.com/@ekrem.hatipoglu/machine-learning-classification-logistic-regression-part-8-b77d2a61aae1

**[10]** Muratlar, E. R. (2020, April 29). *Lojistik regresyon nedir?* Veri Bilimi Okulu: https://www.veribilimiokulu.com/blog/lightgbm/

**[11]** Muratlar, E. R. (2020, March 24). *XGBoost Nasıl Çalışır? Neden İyi Performans Gösterir?* Veri Bilimi Okulu: https://www.veribilimiokulu.com/blog/xgboost-nasil-calisir

**[12]** Şener, Y. (2020, October 25). *Makine Öğrenmesinde Değişken Seçimi (Feature Selection) Yazı Serisi: Genel Bakış*. Medium: https://yigitsener.medium.com/makine-öğrenmesinde-değişken-seçimi-feature-selection-yazı-serisi-genel-bakış-6ac5013d1ee

**[13]** Şengül Gedleç, H. B. (2020, March 2020). *Python Uygulaması ile Karar Ağaçları*. Data Scienc Earth: https://www.datasciencearth.com/python-uygulamasi-ile-karar-agaclari/

**[14]** Willems, K. (2019, December 10). *Keras Tutorial: Deep Learning in Python*. Data Camp: https://www.datacamp.com/community/tutorials/deep-learning-python

**[15]** J. Kruppa, A. Schwarz, G. Arminger, A. Ziegler, Consumer credit risk: Individual probability estimates using machine learning, Expert Syst. Appl. 40 (13) (2013) 5125–5131, http://dx.doi.org/10.1016/j.eswa.2013.03.019.

**[16]** Danenas, P, Garsva, G, (2012), "Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach", Procedia Computer Science, Sayı 9, 1324 – 1333.

**[17]** A. M. (2020, December 1). Kaggle: https://www.kaggle.com/mathchi/credit-risk-evaluation

**[18]** https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).