**A cover of a book

Description automatically generated**

**Table of Contents**

[**Table of Figures** 4](#_Toc184830153)

[**INTRODUCTION** 6](#_Toc184830154)

[DATA PRE-PROCESSING 7](#_Toc184830155)

[UNI VARIATE & BI VARIATE ANALYSIS 9](#_Toc184830156)

[HYPOTHESIS TESTING 12](#_Toc184830157)

[**Hypothesis 1** 12](#_Toc184830158)

[**Hypothesis 2** 12](#_Toc184830159)

[**Hypothesis 3** 13](#_Toc184830160)

[FEATURE ENGINEERING 15](#_Toc184830161)

[PREDICTOR ANALYSIS AND RELEVANCY 15](#_Toc184830162)

[DATA PARTITIONING 17](#_Toc184830163)

[**MODELLING** 17](#_Toc184830164)

[Profit Analysis 20](#_Toc184830165)

[Model Evaluation Comparison 21](#_Toc184830166)

[OBSERVATIONS 21](#_Toc184830167)

[Recommendations 22](#_Toc184830168)

[CONCLUSION 22](#_Toc184830169)

[Executive Summary 23](#_Toc184830170)

[INTRODUCTION 25](#_Toc184830171)

[DATA PRE-PROCESSING 26](#_Toc184830172)

[PREDICTOR ANALYSIS AND RELEVANCY 32](#_Toc184830173)

[DATA TRANSFORMATION 33](#_Toc184830174)

[DATA PARTITIONING 34](#_Toc184830175)

[MODELLING 35](#_Toc184830176)

[Evaluation 35](#_Toc184830177)

[COMPARISON OF MODEL PERFORMANCE 38](#_Toc184830178)

[COST FUNCTION 38](#_Toc184830179)

[PROFIT ANALYSIS 38](#_Toc184830180)

[BUSINESS INTERPRETATION 39](#_Toc184830181)

[OVERALL BUSINESS IMPACT 39](#_Toc184830182)

[RECOMMENDATIONS FOR IMPROVING LOAN DECISIONS 40](#_Toc184830183)

[CONCLUSION 41](#_Toc184830184)

[FUTURE SCOPE 41](#_Toc184830185)

[Executive Summary 42](#_Toc184830186)

[**INTRODUCTION** 44](#_Toc184830187)

[DATA PRE-PROCESSING 45](#_Toc184830188)

[UNI – VARIATE ANALYSIS 48](#_Toc184830189)

[DATA TRANSFORMATION 54](#_Toc184830190)

[DIMENSION REDUCTION 54](#_Toc184830191)

[DATA PARTITIONING 55](#_Toc184830192)

[MODEL SELECTION 55](#_Toc184830193)

[UN-SUPERVISED LEARNING 56](#_Toc184830194)

[PREDICTOR RELEVANCY 59](#_Toc184830195)

[SUPERVISED LEARNING 59](#_Toc184830196)

[CONCLUSION 60](#_Toc184830197)

[RECOMMENDATIONS 61](#_Toc184830198)

[Executive Summary 62](#_Toc184830199)

**Table of Figures**

[Fig1 Department vs Left 1 11](#_Toc184830271)

[Fig2 Salary vs Left 1 11](#_Toc184830272)

[Fig3 Work\_Accident vs Left 1 12](#_Toc184830273)

[Fig4 Promotion in Last 5yrs vs Left 1 12](#_Toc184830274)

[Fig5 DeepDive of Promotion Last 5Years 1 12](#_Toc184830275)

[Fig6 Last Evaluation vs Left 1 12](#_Toc184830276)

[Fig7 Boxplot Satisfaction vs Left 1 13](#_Toc184830277)

[Fig8 Boxplot No.of Projects vs Left 1 13](#_Toc184830278)

[Fig9 Boxplot Time\_Spend vs Left 1 13](#_Toc184830279)

[Fig10 Hypothesis Testing on Salary 1 14](#_Toc184830280)

[Fig11 HypothesisTesting on WorkAccidents 1 14](#_Toc184830281)

[Fig12 Distribution of Promotion 1 15](#_Toc184830282)

[Fig13 Avg Monthly hours Split 1 15](#_Toc184830283)

[Fig14 Distribution of No of Projects 1 16](#_Toc184830284)

[Fig15 Hypothesis3\_GoodPlaceToGrow 1 16](#_Toc184830285)

[Fig16 CorrelationMatrix 1 18](#_Toc184830286)

[Fig17 Varaible\_Selection with LASSO 1 18](#_Toc184830287)

[Fig18 Confusion\_Matrix\_LogisticRegressn 1 20](#_Toc184830288)

[Fig19 Decision\_Tree 1 21](#_Toc184830289)

[Fig20 ConfusionMatrix\_Decision\_Tree 1 21](#_Toc184830290)

[Fig21 Model Evaluation Comparison 1 23](#_Toc184830291)

[Fig22 AUROC Curve 1 23](#_Toc184830292)

[**Fig1 Distribution of the Missing Value 1** 29](#_Toc184830293)

[**Fig2 Distribution Before & After Imputat 1** 31](#_Toc184830294)

[**Figure 3 Reason vs Default 1** 31](#_Toc184830295)

[**Fig 6 Mortgage Due vs Default 1** 32](#_Toc184830296)

[**Fig7 Property vs Default 1** 32](#_Toc184830297)

[**Figure 8 Years on Job vs Default 1** 33](#_Toc184830298)

[**Figure 9 Derog vs Default 1** 34](#_Toc184830299)

[**Figure 10 Delinq vs Default 1** 34](#_Toc184830300)

[**Figure11 Credit Age vs default 1** 34](#_Toc184830301)

[**Figure12 Credit Inquiries vs default 1** 34](#_Toc184830302)

[**Fig13 CreditLines vs Default 1** 35](#_Toc184830303)

[**Figure 14 Correlation Matrix 1** 35](#_Toc184830304)

[**Figure 15 Feature Importance 1** 36](#_Toc184830305)

[**Fig19 Distribution of Imbalanc & Balance 1** 37](#_Toc184830306)

[**Fig21 Logistic\_CostMatrix\_60-40\_split 1** 38](#_Toc184830307)

[**Fig22 Logistic\_ConfusionMatrix\_70-30\_spl** 1 39](#_Toc184830308)

[**Fig23 Logistic\_CostMatrix\_70-30\_split 1** 39](#_Toc184830309)

[**Fig24 Logistic\_ConfusionMatrix\_80-20\_spl 1** 39](#_Toc184830310)

[**Fig25 Logistic\_CostMatrix\_80-20\_split 1** 39](#_Toc184830311)

[**Fig26 Decision Tree 1** 40](#_Toc184830312)

[**Fig27 DecisonTre\_ConfusnMatx\_60-40\_spt 1** 40](#_Toc184830313)

[**Fig28 DecisonTree\_CostMatrix\_60-40 1** 40](#_Toc184830314)

[**Fig1 Distribution of the Missing Value 1** 30](#_Toc184830315)

[**Fig2 Distribution Before & After Imputat 1** 32](#_Toc184830316)

[**Figure 3 Reason vs Default 1** 32](#_Toc184830317)

[**Fig 6 Mortgage Due vs Default 1** 33](#_Toc184830318)

[**Fig7 Property vs Default 1** 33](#_Toc184830319)

[**Figure 8 Years on Job vs Default 1** 33](#_Toc184830320)

[**Figure 9 Derog vs Default 1** 34](#_Toc184830321)

[**Figure 10 Delinq vs Default 1** 34](#_Toc184830322)

[**Figure11 Credit Age vs default 1** 34](#_Toc184830323)

[**Figure12 Credit Inquiries vs default 1** 34](#_Toc184830324)

[**Fig13 CreditLines vs Default 1** 35](#_Toc184830325)

[**Figure 14 Correlation Matrix 1** 35](#_Toc184830326)

[**Figure 15 Feature Importance 1** 36](#_Toc184830327)

[**Fig19 Distribution of Imbalanc & Balance 1** 37](#_Toc184830328)

[**Fig21 Logistic\_CostMatrix\_60-40\_split 1** 38](#_Toc184830329)

[**Fig22 Logistic\_ConfusionMatrix\_70-30\_spl** 1 39](#_Toc184830330)

[**Fig23 Logistic\_CostMatrix\_70-30\_split 1** 39](#_Toc184830331)

[**Fig24 Logistic\_ConfusionMatrix\_80-20\_spl 1** 39](#_Toc184830332)

[**Fig25 Logistic\_CostMatrix\_80-20\_split 1** 39](#_Toc184830333)

[**Fig26 Decision Tree 1** 40](#_Toc184830334)

[**Fig27 DecisonTre\_ConfusnMatx\_60-40\_spt 1** 40](#_Toc184830335)

[**Fig28 DecisonTree\_CostMatrix\_60-40 1** 39](#_Toc184830336)

# **INTRODUCTION**

This project focuses on analyzing a dataset called "Employee.csv" using Human Capital Analytics to identify reasons why employees might leave a company. Earlier, companies used to focus mainly on having the best machines to stay ahead of their competition. But now, the companies that have the most engaged and productive employees are more likely to succeed. That’s why businesses need to keep their valuable employees and understand why some might leave, so they can prevent that from happening.

Initially, we’ll perform descriptive analytics as a foundation for understanding the data. This step involves exploratory data analysis (EDA), which includes techniques like generating summary statistics and visualizations (e.g., histograms, and box plots). These methods help us straightforwardly interpret the data before moving on to more advanced machine-learning techniques for data mining.

**Business Goal**

We aim to understand why employees are leaving our company. By analyzing the factors in the data, we aim to identify key reasons for their departure. This will help us make improvements to keep our top talent and stay competitive. Ultimately, retaining our best employees is crucial for our long-term success and avoiding financial risks from losing them to competitors.

**Cause**

Low job satisfaction and limited career advancement opportunities often lead employees to seek new jobs. Long working hours can cause burnout, while insufficient hours might lead to boredom. These factors contribute to higher employee turnover.

**Effect**

The effects of high employee turnover are quite impactful. Frequent loss of staff leads to substantial costs associated with hiring and training new employees. This constant turnover disrupts work continuity and reduces overall productivity. Moreover, a high turnover rate can lower the morale of the remaining employees, negatively affecting their engagement and work satisfaction.

**Analytics Goal**

The primary objective is to prevent employee attrition by thoroughly analyzing employee data. With a dataset comprising 14,999 observations and 10 variables, we aim to uncover general trends, identify the most influential variables, and detect any unusual patterns. By using key factors or predictor variables such as job satisfaction, average number of hours worked, salary, number of projects handled performance evaluation, etc., The business for Human Capital Analytics aims to build classification machine learning models like logistic regression, decision trees, random forest classifiers, and Ada boosting or bagging to see why important employees are leaving by using variables in the data like job satisfaction, salary number of projects working and average number of hours spent. These solutions can provide the company with recommendations and strategies to improve the working environment, which can also help to retain employees.

# DATA PRE-PROCESSING

**List of Variables**

Here are the list of the variables that we are using for our Analysis

|  |  |
| --- | --- |
| Column Number | Variables |
| 1 | satisfaction\_level |
| 2 | last\_evaluation |
| 3 | number\_project |
| 4 | average\_montly\_hours |
| 5 | time\_spend\_company |
| 6 | work\_accident |
| 7 | left |
| 8 | promotion\_last\_5years |
| 9 | sales |
| 10 | salary |

**Variable Description**

**Satisfaction\_level :** Employee’s level of job satisfaction on a scale from 0.1 to 1.

**Last\_evaluation :** Score of the employee's last performance evaluation, ranging from 0.36 to 1.

**number\_project :** Total number of projects the employee is involved in.

**average\_montly\_hours :** Average number of hours worked per month by the employee.

**time\_spend\_company :** Number of years the employee has spent at the company.

**work\_accident :** Indicates whether the employee has had a work-related accident (0 = no, 1 = yes).

**Left :** Indicates whether the employee has left the company (0 = no, 1 = yes).

**promotion\_last\_5years :** Indicates whether the employee has received a promotion in the last 5 years (0=No,1=yes)

**Sales :** Department where the employee works.

**Salary :** Employee's salary level (e.g., low, medium, high).

**Structure of the Data**

We have a total of 14,999 observations and 10 variables in our dataset, out of 10 variables, 2 are categorical variables, 2 are numerical and 6 are integer variables.

satisfaction\_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...

last\_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...

number\_project : int 2 5 7 5 2 2 6 5 5 2 ...

average\_montly\_hours : int 157 262 272 223 159 153 247 259 224 142 ...

time\_spend\_company : int 3 6 4 5 3 3 4 5 5 3 ...

Work\_accident : int 0 0 0 0 0 0 0 0 0 0 ...

left : int 1 1 1 1 1 1 1 1 1 1 ...

promotion\_last\_5years : int 0 0 0 0 0 0 0 0 0 0 ...

sales : chr "sales" "sales" "sales" "sales" ...

salary : chr "low" "medium" "medium" "low" ...

**Predictors and Target Variable**

Left is our target variable which is an integer with 1 and 0. All other variables like satisfaction level, last evaluation, number of projects, average monthly hours spent in the company, etc., are the predictor variables

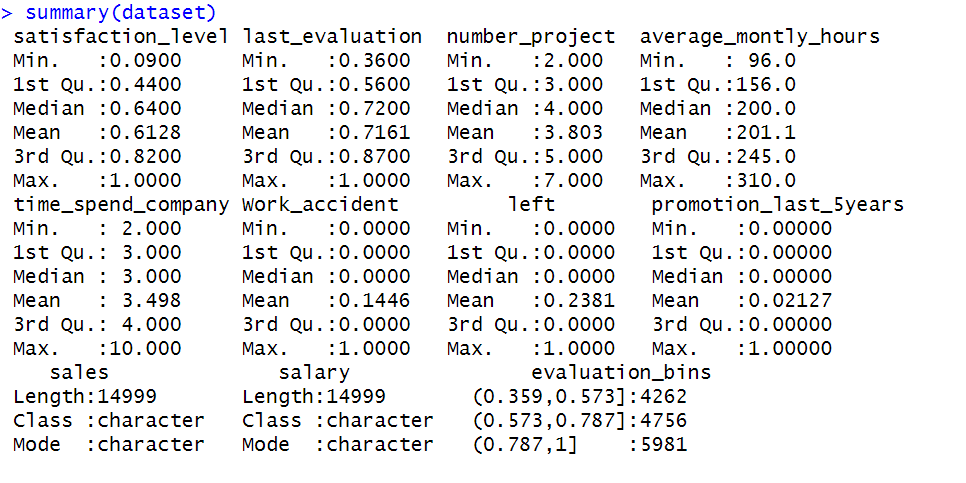
**Checking for the Missing Values and Null Values**

A computer screen shot of a computer code

Description automatically generated

* There are no Missing Values and Null Values in the data.

**Summary of Overall Data**

****

**Summary of Existing (Left=0) Data**

**Total 11,428 Observations (which is 76.1% of Overall data) with 10 variables.**

**A screenshot of a computer screen

Description automatically generated**

**Summary of Churn (Left=1) Data**

**Total 3,571 Observations (which is 23.8% of the overall data) with 10 variables**

A screenshot of a computer

Description automatically generated

**Comparison of Statistics between Existing and Churned Employees from the Company:**

* The average number of projects handled **(“5”)** by churned employees is more than the existing employees which is **“3”**.
* The average monthly hours spent by the churned employees is “**261.2**” which is more compared to existing employees which is **“199.1”**

# UNI VARIATE & BI VARIATE ANALYSIS

**Department vs Left**

A graph of blue and white bars

Description automatically generated

Fig1 Department vs Left 1

The above figure stating that **“Sales”** department employees are more likely to churn compared to other departments followed by Technical and Support.

**Salary vs Left**

A graph of a graph with blue bars

Description automatically generated with medium confidence

From the figure, we can see that Low-Income group employees are more likely to churn followed by medium, this states that income could be one of the reasons for churn

Fig2 Salary vs Left 1

**Work Accident vs Left**

A graph of a work accident

Description automatically generatedEmployees who left have a slightly higher proportion of accidents than those who stayed, suggesting a potential link between work accidents and employee turnover.

Fig3 Work\_Accident vs Left 1

**Promotion in Last 5years vs Left**

A graph of a person's status

Description automatically generatedThe chart shows that most employees, whether they stayed or left, did not receive a promotion in the last 5 years. Promotions are rare and do not appear to have a strong impact on employee retention.

Fig4 Promotion in Last 5yrs vs Left 1

**Promotions Last 5Years vs Left (Deep Dive)**

|  |  |  |
| --- | --- | --- |
|  | No Promotion | Promotion |
| Existing | 11,128 (74.2%) | 300 (2.0%) |
| Churned | 3,557 (23.7%) | 19 (0.1%) |

Fig5 DeepDive of Promotion Last 5Years 1

There are **3,557** employees (**23.7%**) who churned out from the Overall (14,999) who had not been promoted from the last 5 years.

**Last Evaluation vs Left**

A graph of different colored bars

Description automatically generated with medium confidence

From the above figure we can see that Low (0.359 to 0.573) and High Evaluation bins (0.787 to 1) are more likely to churn out.

Fig6 Last Evaluation vs Left 1

**Checking for the Outliers and the Distribution of Data Using Boxplots**

**Satisfaction Level vs Left**

A graph with a box and a line

Description automatically generated with medium confidence

The average satisfaction level of the churned employees is below “**0.50”.** This is meaningful because the employees are less satisfied with working in the company.

Fig7 Boxplot Satisfaction vs Left 1

**Number of Projects Handled vs Left**

A graph of a diagram

Description automatically generated with medium confidence

From the above figure, we can see that there are outliers in the number of projects handling by the employees who are staying in the company which means only very few employees are handling 6 projects, but maximum employees are handling 4 projects. Since our focus is on those who left the company, we can see that the average number of projects handled is 4 and there are no outliers in this data.

Fig8 Boxplot No.of Projects vs Left 1

**Time Spend at Company vs Left**

A graph with a number of rectangular objects

Description automatically generated with medium confidence

The average time spent in the company by the employees who left is “**4”** years and minimum is **“3”** years

Fig9 Boxplot Time\_Spend vs Left 1

# HYPOTHESIS TESTING

**Hypothesis 1** - **Salary (S*alary is the reason why the employees left the company)***

**Null Hypothesis (H0**): An employee’s salary level (`salary`) has no effect on whether they leave (`left`)

**Alternative Hypothesis (H1):** An employee’s salary level significantly affects whether they leave.

|  |  |  |
| --- | --- | --- |
| **Salary** | **Count of salary** | **%Distribution** |
| high | 82 | 2% |
| low | 2172 | 61% |
| medium | 1317 | 37% |

**A graph of a salary level

Description automatically generated**

Fig10 Hypothesis Testing on Salary 1

**Chi-square test**

**A screenshot of a computer code

Description automatically generated**

**Conclusion**

The Chi-square test results for salary show a p-value of less than 0.05 (in fact, much smaller at < 2.2e-16), indicating a statistically significant relationship between salary and whether an employee leaves the company. This means we eject the null hypothesis (H₀)and conclude that salary levels significantly affect employee turnover. Addressing salary concerns could help improve retention rates.

**Hypothesis 2** - **Work Accidents (*Employees leave the company because work is not safe)***

**Null Hypothesis (H0):** Having a work accident (`Work\_Accident`) does not affect whether an employee leaves (`left`).

**Alternative Hypothesis (H1**): Having a work accident significantly affects whether an employee leaves.

|  |  |  |
| --- | --- | --- |
| **Work Accident** | **Count of Work\_accident** | **%Distribution** |
| 0 (No Work Accident) | 3402 | 95% |
| 1 (Yes) | 169 | 5% |

A graph of a number of people

Description automatically generated

Fig11 HypothesisTesting on WorkAccidents 1

**Chi-square test**

**A computer code with blue text

Description automatically generated**

**Conclusion**

We **reject the null hypothesis (H₀)** because the p-value (< 2.2e-16) is significantly less than 0.05. This means that work accidents are significantly associated with employees leaving the company.

**Hypothesis 3** - **Not a Good Place to Grow VS Left (*This company is a good place to grow professionally***

**Null Hypothesis (H0):** This Company is not a good place to grow professionally.

**Alternative Hypothesis (H1):** This Company is a good place to grow professionally

**Note:** Here we are choosing 3 factors like Promotion in Last 5years, Average Monthly hours, Number of Projects Working

**Checking Promotion in the Last 5 Years**

|  |  |  |
| --- | --- | --- |
| **promotion in the last 5 years** | **Count of promotions in the last 5 years** | **%Distribution** |
| 0 (Not Promoted) | 3552 | 99% |
| 1 (Promoted) | 19 | 1% |

A graph with a rectangle

Description automatically generated

Fig12 Distribution of Promotion 1

Form the Above Figure, Promotions are extremely rare, with 99% of employees not receiving one in the last 5 years. This lack of growth opportunities might demotivate employees and lead to higher turnover. Offering more promotions could improve employee satisfaction and retention.

A graph of a graph

Description automatically generated**Check for Average Monthly Hours**

|  |  |  |
| --- | --- | --- |
| **Hours Bucket** | **Count of left** | **%Distribution** |
| 126-205 | 1662 | 47% |
| 206-285 | 1540 | 43% |
| 286-365 | 369 | 10% |

Fig13 Avg Monthly hours Split 1

Most employees who left worked between 126-205 hours per month (47%), followed by 206-285 hours (43%). Only 10% worked 286-365 hours, suggesting that very low or very high work hours may be linked to employee turnover.

|  |  |  |
| --- | --- | --- |
| **No.of Projects** | **Count of number\_project** | **%Distribution** |
| 2 | 1567 | 44% |
| 3 | 72 | 2% |
| 4 | 409 | 11% |
| 5 | 612 | 17% |
| 6 | 655 | 18% |
| 7 | 256 | 7% |

**Check for Number of Projects**

A graph of a bar graph

Description automatically generated

Fig14 Distribution of No of Projects 1

Most employees (44%) worked on just 2 projects, while the number of employees decreases as the number of projects increases. A significant portion worked on 5-6 projects (35%), which could indicate a workload sweet spot, while fewer employees handled very low (3 projects) or very high (7 projects) workloads.

**Final Analysis**

From the references taken from the above, like average working hours, no promotion, and number of projects worked, we are considering the subset of churned employees who have not been promoted in the last 5 years with evaluation of high and medium bins. They work more than the average hours, which is >=201 per week, and they work on more projects than the average, which is 4, but with a low satisfaction level.

|  |  |
| --- | --- |
| **Left vs not good place to work** | **Count** |
| No of Left | 3571 |
| Not good Professionally | 898 |
|  | 25% |

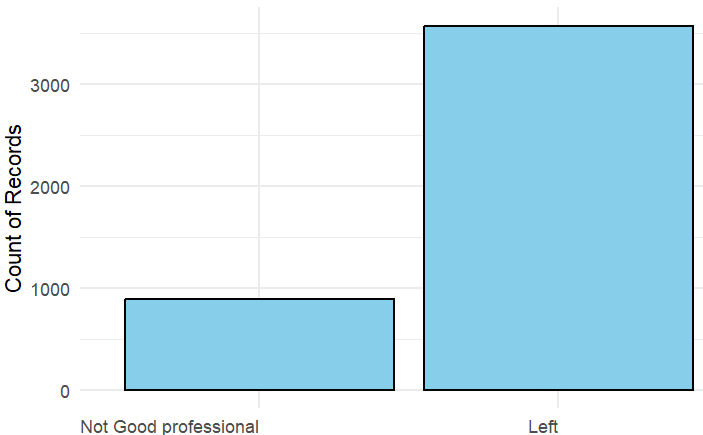


Fig15 Hypothesis3\_GoodPlaceToGrow 1

**Conclusion**

We reject the null hypothesisbecause 25% (898 employees) of churned employees left due to factors like lack of promotion, working more than average hours, handling more projects than average, and low satisfaction levels, the remaining 75% (2673 employees) left for reasons unrelated to these factors. This suggests that despite the challenges for a subset, the workplace offers good professional growth opportunities for most employees.

# FEATURE ENGINEERING

Based on min, max, and mean values of a few variables like time\_spend\_in\_the\_company, last\_evaluation, satisfaction\_level, number\_of\_projects\_worked, had created bins with a normal distribution.

This will help us to simplify the data, easy interpretability and handle the outliers.

A screenshot of a computer screen

Description automatically generated

The above image is the summary of modified dataset with newly created bins like time\_spend\_company\_bucket, number\_project\_bucket and last\_evaluation\_bucket.

**One hot encoding**

One-hot encoding converts categorical variables into a format for machine learning by creating separate columns for each category. Each row gets a '1' in the column corresponding to its category and '0's in all other columns. For example, 'sales', 'HR', and 'accounting' become separate columns, with a '1' indicating the job role and '0's elsewhere. This makes it easier for algorithms to process the data.

If we have a list of job roles like 'sales', 'accounting', and 'HR'. Instead of keeping these job roles as text, which computers can't easily work with, one-hot encoding turns them into numbers. Each job role gets its column, and if someone is in 'sales', they get a '1' in the 'sales' column and a '0' in the others. If they are in 'HR', they get a '1' in the 'HR' column and a '0' in the others.

# PREDICTOR ANALYSIS AND RELEVANCY

**Dimension Reduction**

Dimension reduction simplifies a model by reducing the number of variables, but it might not be needed for our dataset with just 10 variables and 15,000 observations. With only 10 variables, your model can handle the data efficiently without becoming overly complex. The large dataset ensures that the model can learn effectively from each variable.

Dimension reduction is more useful when there are many variables, potentially causing complexity or inefficiencies. Since you have a manageable number of variables, each one contributes valuable information. Keeping all 10 variables helps maintain the full scope of insights and ensures more accurate results.

**Correlation Matrix**

**A diagram of a graph

Description automatically generated with medium confidence**

* There is no strongly correlated (> 0.7) variables in this data.
* Satisfaction level and number of projects which is -0.14 followed by Satisfaction level and time\_spend\_company are weekly correlated with -0.39
* Time\_Spend\_company, Average\_monthly\_hours, number\_project, last\_evaluation showing the positive correlation with the target variable left whereas promotion\_last\_5years, satisfaction\_level, work\_accident is showing negative correlation with the left variable
* Out of all time\_spend\_company is showing a more positive correlation with left with 0.14 followed by average\_monthly\_hours which is 0.07

Fig16 CorrelationMatrix 1

**Lasso Regression**

Lasso regression is a type of linear regression that helps prevent overfitting by adding a penalty to the model for having too many variables. It forces some less important variables to have coefficients of zero, effectively removing them from the model. This makes the model simpler and easier to interpret. Lasso is especially useful when you have a lot of features and want to identify only the most important ones for making predictions.After applying lasso regression to the dataset.

A graph with different colored bars

Description automatically generated

Fig17 Varaible\_Selection with LASSO 1

Salary, average\_monthly\_hours, sales, time\_spend\_company\_bucket, last\_evaluation and number\_project\_bucket seem like mostly positive variables chosen by lasso regression.

Overall, it seems like all the variables in the dataset are equally important. So, I am considering all the variables to incorporate in my model

# DATA PARTITIONING

Data partitioning is an essential step in machine learning, involving the division of a dataset into training, validation, and testing sets to effectively evaluate model performance.

**Train Data**

This subset of the data is used to train the model. It consists of known input-output pairs, allowing the model to learn and identify patterns and relationships between features (input variables) and the target (output variable). The model uses this data to adjust its parameters and improve its accuracy in predicting outcomes.

**Test Data**

To evaluate how well the model performs on new, unseen data, we use the testing set. This dataset, which the model has not encountered during training, is crucial for assessing generalization. Comparing the model's predictions with the actual values from the testing set helps in identifying any issues, such as overfitting or underfitting, and provides a final measure of the model's performance.

After all, data cleaning, analysis of variables, feature engineering, choosing of relevant variables. The next step is to partition the data into Train, Test, and Validation. There are many methods to partition the data, here I have chosen 60,40.

Here, we have **chosen 60%,40% rules for Train and Test**

A computer screen shot of a computer code

Description automatically generated

The above figure explains that there are 8,661 observations in train dataset, 2,887 observations in the test dataset, and 2,887 observations in the validation dataset.

# **MODELLING**

Since our target is the **“left”** variable which has outcomes of ‘0’ and ‘1’.

**Supervised machine learning techniques**

* logistic regression
* Decision Tree
* Gradient boosting

**Logistic Regression**

Logistic regression is a statistical method used to predict a yes/no outcome based on one or more factors. It calculates the probability of a certain event happening and converts this probability into a 0 or 1 outcome. This technique is popular because it's easy to understand, efficient, and can provide clear results.

A screenshot of a computer code

Description automatically generated

Logistic Regression by itself have a capability to choose the signification variables, here we can see a highly significant variable with 3stars(\*\*\*) with p-value less than < 0.05. example: Satisfaction\_level,number\_project, average\_monthly\_hours etc.,

**Confusion Matrix**

**A screenshot of a computer

Description automatically generated**

**Observations**

The Logistic Regression has predicted 36.75% of churned employees(Sensitivity) and 92.96% of existing employees (Specificity) with an accuracy of 79.53%. and error rate of 21%

Fig18 Confusion\_Matrix\_LogisticRegressn 1

**Decision Tree**

A decision tree is a simple method for making predictions by splitting data into smaller parts based on certain rules. It works like a flowchart, where each step is a decision based on a feature, and the outcome is predicted. Decision trees can work with both numbers and categories, making them flexible and easy to understand. However, they can make too-specific predictions if not carefully controlled, leading to overfitting.

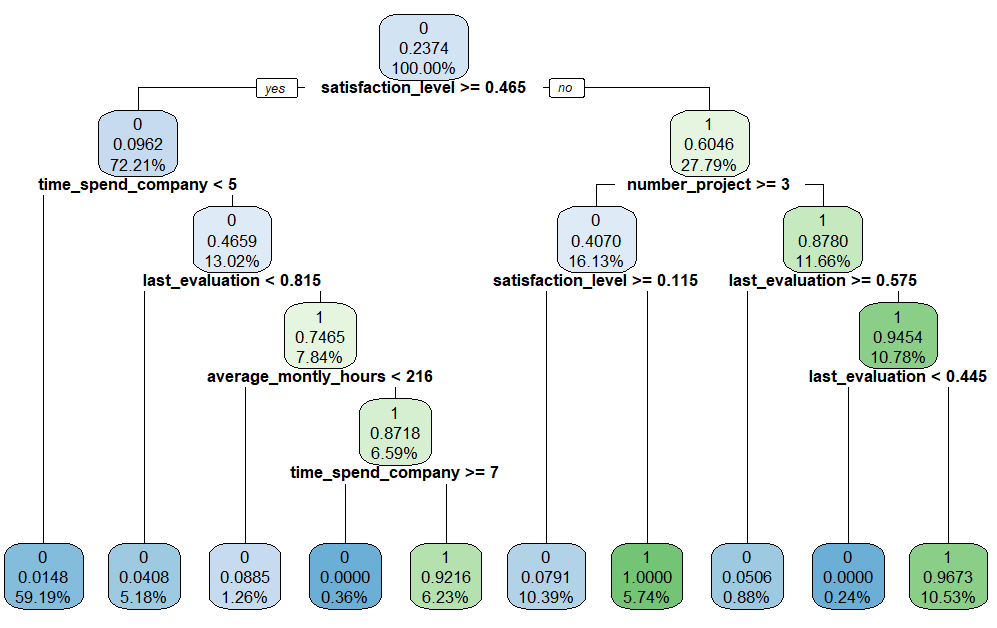


Fig19 Decision\_Tree 1

**Interpretation**

**Rule 1:**

If `satisfaction\_level >= 0.465` and `time\_spend\_company < 4.5`

If an employee's satisfaction level is **0.465 or higher** and they have been with the company for **less than 4.5 years**, they are very likely to **stay**.

In this case, 98.5% of employees in this group stay at the company.

**Rule 2:**

If `satisfaction\_level >= 0.465` and `time\_spend\_company >= 4.5` and `last\_evaluation >= 0.815` and `average\_monthly\_hours >= 215.5` and `time\_spend\_company < 6.5`

If an employee's satisfaction level is **0.465 or higher**, they have been with the company for **4.5 years or more**, their last evaluation score is **0.815 or higher**, and they work **more than 215.5 hours per month**, but they have been with the company for **less than 6.5 years**, they are likely to **leave**.

Among these employees, 92.2% leave the company.

**Rule 3:**

If `satisfaction\_level < 0.465` and `number\_project < 2.5` and `last\_evaluation < 0.575` and `last\_evaluation >= 0.445`

Then the predicted class is `1` (leave), with 96.7% confidence (96.7% of the data points in this node are classified as `1`)

**Confusion Matrix**

**A screenshot of a computer code

Description automatically generated**

**Observations**

The decision tree has predicted 91% of churned employees(Sensitivity) and 98% of existing employees (Specificity) with accuracy of 97% with an error rate of 3%

Fig20 ConfusionMatrix\_Decision\_Tree 1

# Profit Analysis

Based on our best performing model Decision Tree with 60-40 split, we are checking which group of confusion matrix belongs to which salary salary group

A computer screen shot of a computer code

Description automatically generated

**Key Definitions**

TP (True Positives): Employees correctly predicted to leave the company.

TN (True Negatives): Employees correctly predicted to stay at the company.

FP (False Positives): Employees incorrectly predicted to leave but stayed.

FN (False Negatives): Employees incorrectly predicted to stay but actually left.

**Breakdown by Salary Groups**

**1. High Salary**

* **FN**: 4 employees were incorrectly predicted to stay but actually left.
* **FP**: 6 employees were incorrectly predicted to leave but stayed.
* **TN**: 459 employees were correctly predicted to stay.
* **TP**: 25 employees were correctly predicted to leave.

**2. Low Salary**

* **FN**: 76 employees incorrectly predicted to stay but actually left.
* **FP**: 19 employees incorrectly predicted to leave but stayed.
* **TN**: 2,016 employees correctly predicted to stay.
* **TP**: 812 employees correctly predicted to leave.

**3. Medium Salary**

* **FN**: 42 employees incorrectly predicted to stay but actually left.
* **FP**: 24 employees incorrectly predicted to leave but stayed.
* **TN**: 2,041 employees correctly predicted to stay.
* **TP**: 475 employees correctly predicted to leave.

1. **Churn Prediction Effectiveness**:
   * For the **low salary** group, the highest number of employees (812) were correctly identified as leaving the company, reflecting the model's strength in predicting churn for this category.
   * For the **medium salary** group, 475 employees were correctly predicted to leave, with fewer errors compared to the low salary group.
   * For the **high salary** group, only 25 employees were correctly predicted to leave, reflecting a smaller sample size for this category.
2. **Error Analysis**:
   * **FN**: Low salary employees had the highest false negatives (76), meaning many employees who left were incorrectly predicted to stay. This may indicate the model struggles with predicting churn in this group accurately.
   * **FP**: False positives were relatively low across all salary groups, indicating fewer employees were incorrectly predicted to leave.

**Overall**

* The model performs well across all salary groups, with high accuracy (97-98%) in predicting whether employees will stay or leave.
* The low salary group had the largest number of correctly identified churned employees (812), but it also had the highest errors in false negatives (76), meaning many low-salary employees who left were missed by the model.
* The high salary group had the least churn overall, reflecting better retention for high earners, but its small size limits insight into errors.
* Overall, focusing on reducing errors (especially false negatives) in the low and medium salary groups could further improve retention strategies.

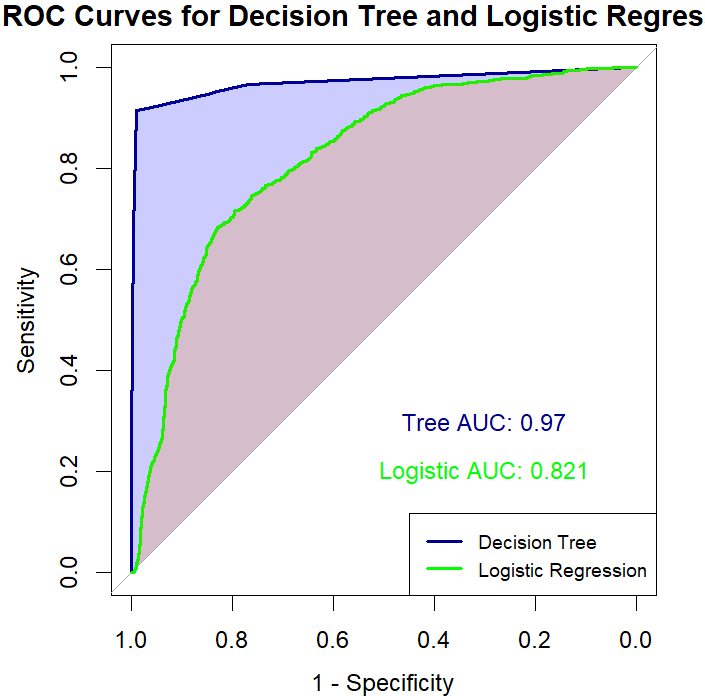
# Model Evaluation Comparison

From the figure below, we can see that the decision tree with a split of 60% and 40% is performing well compared to other models in terms of accuracy, which is 97%. Although it has low sensitivity, it has the high sensitivity of 91% and low error rate of 3%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **`** | **Accuracy** | **Sensitivity** | **Specificity** | **Error Rate** |
| **Logistic Regression** | 79.53% | 36.75% | 92.96% | 20.47 |
| **Decision Tree** | 97.15% | 91.49% | 98.93% | 2.85 |

Fig21 Model Evaluation Comparison 1

**AUROC curve**

The ROC curve helps us evaluate how well a classification model performs by plotting its true positive rate against its false positive rate. It shows how the model's accuracy changes with different decision thresholds. The AUC (Area Under the Curve) quantifies the overall ability of the model to distinguish between the two classes. A higher AUC value indicates better model performance, with 1 representing perfect accuracy and 0.5 indicating no better than random guessing.

From the above graph we can say that the Decision Tree has performed well with AUC of 0.97, indicating that the model has balanced the trade-off between detecting positives (True Positives Rate) and avoiding false positives (False Positive Rate). This high value shows that the model performs strongly despite our chosen threshold.

Fig22 AUROC Curve 1

# OBSERVATIONS

* Based on our hypothesis, out of 3571 who left the company, 25% of employees worked more than average no of projects, good evaluation, with no promotion have a low satisfaction level, since it is 25% we cannot say that this is not a good place to work
* Salary seems one of the reasons for churn from the chi-square test which has a p-value <0.05 but work accident is not a reason for leave.

# Recommendations

**Reason 1: Low Satisfaction and Long Tenure**

**Condition**: Employees with `satisfaction\_level < 0.465` and `time\_spend\_company >= 4.5 years`

**Why They Might Leave:** Employees who have been with the company for a long time and are dissatisfied are more likely to consider leaving. They might feel stuck or unappreciated despite their years of service.

**Recommendation**

Increase job satisfaction by recognizing their contributions, providing new challenges, and offering career advancement opportunities. Regularly check in to address any concerns they might have.

**Reason 2: High Workload and Moderate Satisfaction**

**Condition:** Employees with `satisfaction\_level >= 0.465`, `time\_spend\_company >= 4.5 years`, `last\_evaluation >= 0.815`, and `average\_monthly\_hours >= 215.5` but `time\_spend\_company < 6.5 years`

**Why They Might Leave:** Employees with high satisfaction but facing high workloads and long hours may feel overwhelmed or burned out, leading them to consider leaving.

**Recommendation**

Manage workloads more effectively. Consider redistributing tasks, providing additional support, or ensuring that employees have adequate breaks and work-life balance.

**Reason 3: Low Satisfaction, Few Projects, and Poor Evaluation**

**Condition:** Employees with `satisfaction\_level < 0.465`, `number\_project < 2.5`, and `last\_evaluation between 0.445 and 0.575`

Why They Might Leave: Dissatisfied Employees, working on fewer projects, and have mediocre evaluations may feel underutilized or undervalued, leading to a higher likelihood of leaving.

**Recommendation**

Increase their involvement in more projects to boost engagement and provide training to improve their performance. Address dissatisfaction by creating a more supportive work environment and offering constructive feedback.

# CONCLUSION

Employees are likely to leave if they feel unhappy and have been with the company for a long time without new opportunities. High workloads can also lead to burnout, causing dissatisfaction. Additionally, employees who feel underutilized and have low-performance evaluations may decide to leave. To retain employees, it's crucial to improve job satisfaction, balance workloads, and offer meaningful work and growth opportunities.

Name: Priyanka Gonugunta

Project: Human Capital Analytics to Improve Employee Retention

# Executive Summary

**Objective**

This project investigates the factors influencing employee retention and turnover within a company, utilizing a dataset from Human Capital Analytics. The analysis focuses on key variables such as job satisfaction, promotion history, work accidents, and time spent at the company. By applying descriptive analytics and exploratory data analysis, the project aims to identify patterns and outliers that may indicate why employees choose to leave.

**Approach**

The findings reveal significant relationships between employee turnover and various factors, particularly job satisfaction and promotion frequency. Visualizations like stacked bar charts and correlation analysis help clarify these relationships. The insights gained from this analysis can guide companies in implementing targeted strategies to improve employee retention, thereby reducing turnover and maintaining a competitive advantage in the market.

# INTRODUCTION

This project focuses on analyzing a dataset known as "Loan Data.csv" using data analytics to understand the factors influencing loan defaults. Traditionally, financial institutions prioritized maximizing profits through high loan volumes, often neglecting the significance of understanding borrower behavior. In today’s competitive landscape, organizations that leverage data to assess risks and identify potential defaults are better positioned to succeed. Therefore, it’s crucial to analyze why certain borrowers default on their loans, allowing lenders to develop proactive strategies to mitigate risks.

We will begin by conducting descriptive analytics as a foundation for understanding the dataset. This phase involves exploratory data analysis (EDA), which includes generating summary statistics and visualizations (such as histograms and correlation matrices). These techniques will provide clear insights into the data, setting the stage for applying advanced machine learning techniques aimed at predicting loan defaults and identifying key risk factors.

**Business Problem**

The business objective is to help the bank predict which customers are at risk of defaulting on their loans. By identifying these high-risk customers early, the bank can take preventative measures, such as adjusting loan terms or offering financial support, to reduce defaults and protect its financial health. This approach will also lead to better loan approval decisions and improved risk management overall.

**Cause**

The bank is using a non-intelligent, traditional approach for making loan decisions, without considering detailed data-driven risk assessments. So, evaluating the risk will become very tough and chances of grating the loans to the high-risk individuals will increase.

**Effect**

As a result of this non-intelligent approach, the bank is losing money due to loan defaults, as it is unable to effectively identify high-risk borrowers early in the process. This impacts on the bank's financial health, leading to potential revenue losses.

**Business Goal**

The objective is to help the bank predict which customers are likely to default on their loans, reducing financial losses and improving decision-making in loan approvals.

**Analytics Goal**

The primary objective is to prevent employee attrition by thoroughly analyzing Loan data. With a dataset comprising 5,961 observations and 13 variables, we aim to uncover general trends, identify the most influential variables, and detect any unusual patterns. The goal is to build a predictive model that identifies key factors influencing loan defaults and accurately predicts which applicants are most likely to default. This will enable the bank to assess credit risk more effectively and make data-driven loan approval decisions.

**Analytics Approach**

The analytical approach involves using data analysis techniques to explore the loan dataset with data exploration, visualization, and basic statistics, then moves on to predictive modeling using machine learning to forecast which customers are likely to default. This will help the bank make more informed decisions and reduce financial risk.

# DATA PRE-PROCESSING

**VARIABLE DESCRIPTION**

**BAD:** Binary indicator (1 or 0) representing whether an applicant defaulted on their loan (1 = default, 0 = no default).

**LOAN:** The total amount of the loan requested by the applicant, measured in monetary units (e.g., dollars).

**MORTDUE:** The outstanding mortgage balance owed by the applicant, expressed in monetary units.

**VALUE:** The current market value of the property used as collateral for the loan, measured in monetary units.

**REASON**: The purpose of the loan is typically categorized (e.g., "Home Improvement").

**JOB:** The applicant's job type or occupation, described as a categorical variable (e.g., "Other," "Professional," etc.).

**YOJ:** Years on the job, indicating how long the applicant has been employed in their current position.

**DEROG:** The number of derogatory marks on the applicant's credit report, such as bankruptcies or foreclosures.

**DELINQ:** The number of times the applicant has been late on payments, indicating their payment history.

**CLAGE:** The age of the oldest credit line the applicant has measured in months.

**NINQ:** The number of recent inquiries into the applicant's credit report, which can reflect their credit-seeking behavior.

**CLNO:** The total number of credit lines the applicant has open, reflecting their credit utilization.

**DEBTINC:** The debt-to-income ratio, calculated by dividing the applicant's total monthly debt payments by their gross monthly income, indicates their ability to manage debt.

**Renaming the Variables**

In this section, we rename the following variables to make them more understandable and meaningful.

* **“BAD”** variable into “**Default”** variable
* **“NINQ”** variable into “**Credit\_Inquiries”** variable
* **“CLNO”** variable into “**Credit\_Lines”** variable

**Structure of the Data**

We have a total of **5,961** observations and **13** variables, out of 13 variables, 2 are categorical variables, 5 are numerical and 6 are integer variables.

**DEFAULT :** int 1 1 1 1 0 1 1 1 1 1 ...

**LOAN :** int 1100 1300 1500 1700 1700 1800 1800 2000 ...

**MORTDUE :** num 25860 70053 13500 NA 97800 ...

**VALUE :** num 39025 68400 16700 NA 112000 ...

**REASON :** chr "HomeImp" "HomeImp" "HomeImp" "" ...

**JOB :** chr "Other" "Other" "Other” " ...

**YOJ :** num 10.5 7 4 NA 3 9 5 11 3 16 ...

**DEROG :** int 0 0 0 NA 0 0 3 0 0 0 ...

**DELINQ :** int 0 2 0 NA 0 0 2 0 2 0 ...

**CLAGE :** num 94.4 121.8 149.5 NA 93.3 ...

**Credit\_Inquiries :** int 1 0 1 NA 0 1 1 0 1 0 ...

**Credit\_Lines :** int 9 14 10 NA 14 8 17 8 12 13 ...

**DEBTINC :** num NA NA NA NA NA ...

**Predictors and Target Variable**

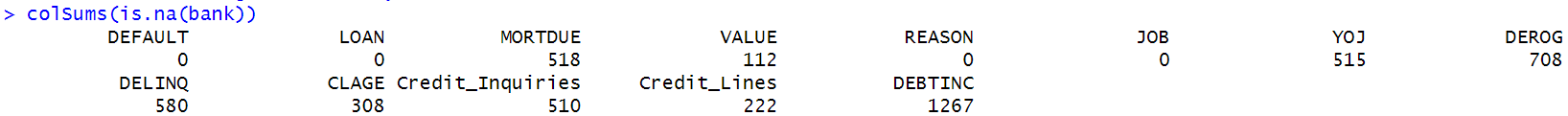
Since our project goal is to classify the default customers. So, our **target** variable will be **Default**, and all other variables like LOAN, REASON, MORTDUE, DELINQ, DEORG, VALUE, JOB, YOJ, CLAGE, Credit\_Inquiries, Credit\_Lines, DEBTINC are predictors.

**MISSING VALUE ANALYSIS**

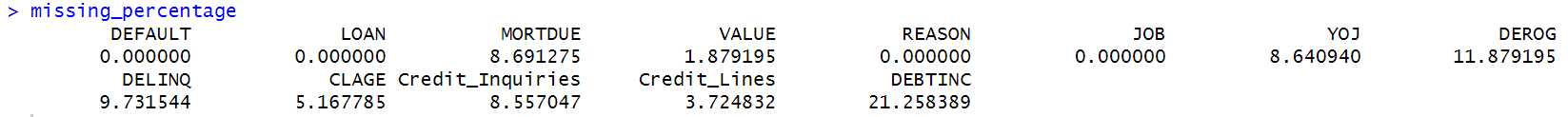
**Checking for the Missing Values and Null Values**

There is a total of 4,740 missing values in the overall data.

**Missing values by columns**

****

**Missing values in Percentage**

****

DEBTINC has the highest percentage of Missing values is 21% followed by DEORG is 12%, DELINQ which is 10% and MORTDUE, YOJ, and Credit Enquiries have 9% and other variables have less than 4% and some variables like BAD, LOAN, REASON, JOB have no Missing Values.

**Distribution of the Missing Value**

From the below figure ,We can see that DEBTINC has the highest number of Missing Values followed by DEROG and DELINQ.

**A graph of red bars

Description automatically generated with medium confidence**

**Fig1 Distribution of the Missing Value 1**

**Excluding Observations with more than 50% of columns empty**

These are 164 observations, which have more than 50% of empty rows, meaning more than 7 columns don’t have any data. If the entry of these values happened by mistake. If we try to impute all these observations, it may create biases in our analysis.

**A screenshot of a computer

Description automatically generated**

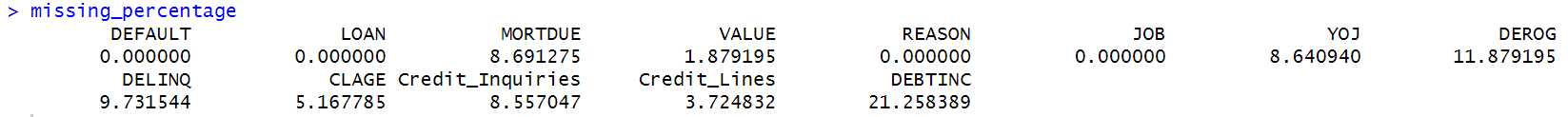
Removing this data will lose 2.7% of data from the analysis (out of 5960, 164 is removed).

**Removing the observations from the DEBTINC column which are Empty**

Since there are **21%** missing values in the **DEBTINC** ratio, we are not including the observations in our analysis.

Because with this range, the impact of imputation could be more pronounced, missingness is related to observed data but not the missing data itself (Missing at Random - MAR).

All other variables which have missing values are less than 15%.

****

**A white background with blue text

Description automatically generated**

**Check for Emptiness in the Categorical Columns**

There are 151 empty records in the REASON variable and 183 empty records in the JOB variable.

**A computer code with blue text

Description automatically generated**

Imputing these empty values with UNKNOWN values. So, it is easy to handle and understandable. Here is the following figure showing how it got replaced in the dataset.

A computer code with blue text

Description automatically generated

**How to handle Missing values?**

We assumed that KNN (K-Nearest Neighbors) would be an effective method for imputing missing values in our dataset. This is because KNN imputes values based on the characteristics of the nearest neighbors in the feature space, which helps maintain the distribution of the data and reduces the likelihood of introducing biases. By considering the similarities between data points, KNN can provide more contextually relevant imputations compared to simpler methods such as mean or median imputation.

A screenshot of a computer code

Description automatically generated

After applying KNN, we can see that there are no missing values in the data.

**Checking for the Distribution of variables before and after Imputation** A screenshot of a graph

Description automatically generated

**Fig2 Distribution Before & After Imputat 1**

We can also see that the data follows a similar distribution after the imputation.

**Uni variate & Bi Variate Analysis on Predictors**

From the above graph, we can see that the data is right-skewed, which means we have a lot of outliers in our data.

**Reason vs Default**

A graph of a number of people

Description automatically generated with medium confidence

From the fig 4 , we can say that all the Reasons are equally distributed with the default variable.

**Figure 3 Reason vs Default 1**

**Job Type vs Default**A graph of different colored columns

Description automatically generated

**Figure 4 JobType vs Default 1**

From the fig 5 graph, we can say the customers who belong to the Sales Job type are more likely to default compared to other Job Types which is followed by the Job type

**Loan vs Default**

**A graph of a diagram

Description automatically generated with medium confidence**

The distribution looks similar in both defaulters and non-defaulters. We can see a few customers who have taken more loans than the average.

**Fig5 Loan vs Default 1**

A graph with lines and numbers

Description automatically generated with medium confidence**Mortgage Due vs Default**

The distribution looks similar in both defaulters and non-defaulters. We can see some customers have taken more mortgage dues than the average number of customers

**Fig 6 Mortgage Due vs Default 1**

**A graph of a property value

Description automatically generatedProperty vs Default**

**Fig7 Property vs default 1**

**Figure7 Property vs Default 1**

The distribution looks similar in both defaulters and non-defaulters. We can see, few customers have more Properties than the average number of customers.

**Fig7 Property vs Default 1**

**Years on Job vs Default**

**Figure 8 Years on Job vs Default 1**

**Fig 8 Years on Job vs Default 1**

A graph with a number of lines

Description automatically generated with medium confidence

The distribution looks similar in both defaulters and non-defaulters. But we can have fewer customers and have more work experience than the average.

**Derog vs Default**

A graph with lines and dots

Description automatically generated

The distribution is skewed to the right side which means more clients are less than the average derogator (not paying the loan on due time), and few customers are outliers, which means fewer of the customers stand more than the average when it comes to missing the payment on time.

**Figure 9 Derog vs Default 1**

**Fig 9 Derog vs Default 1**

**DELINQ vs Default**

A graph with a number of lines

Description automatically generated with medium confidence

**Figure 10 Delinq vs Default 1**

This also shows a similar pattern to Derog. which means distribution is skewed to the right side which means more clients are less than the average delinquency (missing to pay the loan), and few customers are outliers, which means fewer of the customers stand more than the average when it comes to missing the payment on time.

**Credit Age vs Default**

A graph with lines and dots

Description automatically generated

**Figure11 Credit Age vs default 1**

The distribution looks similar in both defaulters and non-defaulters. But we can see few of them have a credit age more than the average.

**Credit Inquiries vs Default**

**Figure12 Credit Inquiries vs default 1**

A graph with lines and dots

Description automatically generated

The distribution looks similar in both defaulters and non-defaulters. But we can see customers who have **Credit Inquiries** more than the average.

**Credit Lines vs Default**

**A graph with a number of credit lines

Description automatically generated**

The distribution looks similar in both defaulters and non-defaulters. But we can see, a few of the customers have **Credit Lines** more than the average.

**Fig13 CreditLines vs Default 1**

# PREDICTOR ANALYSIS AND RELEVANCY

**A diagram of a credit line

Description automatically generatedCorrelation Matrix**

**Positive Correlation**

* There is an 80% positive correlation between Value and Mortdue which is the highest correlation value among all the predictors.
* There is 28% positive correlation between Mortdue and Credit\_lines followed by 27% between loan and value.
* There is a 27% positive correlation between CLAGE and Credit\_Lines.

**Negative Correlation**

* Derog and CLage have negative correlations which is -0.18
* MortDue and YOJ have a negative correlation which is -0.05

**Figure 14 Correlation Matrix 1**

**Correlation with Target Variable Default**

* Except Delinq(0.23), Derog(0.23), and Credit\_Enquiries(0.16) all have negative correlation with the Default Variable. However, these do not state that these variables are strongly correlated except Value and MortDue.

Seems like all variables are important for analysis.

**Boruta Algorithm**

The Boruta algorithm, which assesses feature importance by comparing each feature to randomized counterparts, also confirmed that all features were important. After running 10 iterations, Boruta identified 11 attributes as crucial (Excluding Target variable): `CLAGE`, `Credit\_Inquiries`, `Credit\_Lines`, `DELINQ`, `DEORG`, `LOAN`, and `MORTDUE` etc., No attributes were deemed unimportant,highlighting that every feature in the dataset plays a significant role in the model. This result aligns with the findings from the stepwise regression methods, reinforcing the value of including all predictors in the analysis.

A graph with text and numbers

Description automatically generated with medium confidence

**Figure 15 Feature Importance 1**

A close up of a computer screen

Description automatically generated

From Boruta we can see that all the 12 variables are important for analysis, DEBTINC is most important compared to others.

# DATA TRANSFORMATION

**Convert to Factors**

Here we are converting REASON, JOB, DEROG and DELINQ to factor variables.

**A group of numbers and symbols

Description automatically generated**

**Dimension Reduction**

* Dimension reduction simplifies a model by reducing the number of variables, but it might not be needed for our dataset with just **12** variables and **4,640** observations. With only 12 variables, our model can handle the data efficiently without becoming overly complex. The large dataset ensures that the model can learn effectively from each variable. (Excluding Target Variable).
* Dimension reduction is more useful when there are many variables, potentially causing complexity or inefficiencies. Since you have a manageable number of variables, each one contributes valuable information. Keeping all 12 variables helps maintain the full scope of insights and ensures more accurate results.

# DATA PARTITIONING

After cleaning the data, analyzing variables, and performing feature engineering, the next step is partitioning the data for model training and testing. I used various partitioning methods, including 60-40, 70-30, and 80-20 splits.

Each split is designed to balance training the model with enough data and testing its performance on unseen data. These different splits allow for experimentation to find the optimal division that yields the best model performance.

**A computer screen shot of a number

Description automatically generated**

**Check for the Imbalance in the Dataset**

After Imputing the missing values and removing the unnecessary data, we are left with 4,640 observations and 13 variables including our Target variable Default.

A close-up of a number

Description automatically generatedWe observe that only 8.6% of the data belongs to the minority class (1), while the majority class (0) accounts for 91.3%. Addressing this imbalance is crucial to mitigate biases introduced by the predominant majority class

* **Impact of Imbalance:**
  + Models trained on imbalanced data often achieve high accuracy by predicting the majority class for most observations, neglecting the minority class entirely.
  + For this project, misclassifying defaulters (False Negatives) could result in substantial financial losses for the bank, while overestimating defaults (False Positives) could lead to missed opportunities and customer dissatisfaction.

**How to Handle the Imbalance?**

Employed the Random Over Sampling Examples (ROSE) technique on the training data to create a balanced dataset by generating synthetic samples for the minority class.

ROSE was chosen over simple oversampling to avoid exact duplication of minority class samples, which can lead to overfitting.

A graph of a bar and a bar of a graph

Description automatically generated with medium confidence

**Fig19 Distribution of Imbalanc & Balance 1**

Before applying ROSE, the minority class accounted for only 8.6% of the training data. After applying ROSE, the dataset achieved a near-equal distribution between defaulters and non-defaulters.

The distribution plot of the dataset before and after applying ROSE highlighted the improvement in class balance, enabling the model to learn effectively from both classes.

# MODELLING

Since our target is the **“Default”** variable which has outcomes of ‘0’ and ‘1’.

**Supervised machine learning techniques**

* logistic regression with different partitioning methods
* Decision Tree

**Logistic Regression**

**60- 40 Split**

Here, we are applying logistic regression as a baseline model on our balanced data with 60% as Train and 40% as Test data.

**A computer screen shot of a code

Description automatically generated**

# Evaluation

**Confusion Matrix**

**Logistic Regression with (60 – 40) Split**

**A screenshot of a computer

Description automatically generated**

**Interpretation**

Accuracy is 78%

Sensitivity is 56.66%

Specificity is 80.18%

Error Rate: 1- Accuracy which is 22%

**Fig22 Logistic\_ConfusionMatrix\_60-40\_spl 1**

**Cost Matrix**

**A computer screen shot of a number

Description automatically generated**

* It wrongly identified **234 loans** as likely to default when they wouldn’t, leading to unnecessary concern and possibly lost business opportunities.
* It also missed **46 loans** that did default, which could result in financial losses for the bank.

**Fig21 Logistic\_CostMatrix\_60-40\_split 1**

**Logistic Regression with (70 – 30) Split**

Change the partition size to 70% train and 30% test.

**Confusion Matrix**

A screenshot of a computer

Description automatically generated

**Interpretation**

Accuracy is 78%

Sensitivity is 56.66%

Specificity is 80.18%

Error Rate: 1- Accuracy which is 22%

**Fig22 Logistic\_ConfusionMatrix\_70-30\_spl** 1

**Cost Matrix**

* **A close-up of a computer screen

  Description automatically generated**The model incorrectly classified **252 loans as likely to default**, even though they wouldn’t. This misclassification could cause unnecessary concern for borrowers and might result in lost business opportunities for the bank.
* Additionally, it **missed 52 loans that did default**, leading to potential financial losses for the bank, as these high-risk customers were not identified for additional scrutiny or intervention.

**Fig23 Logistic\_CostMatrix\_70-30\_split 1**

**Logistic Regression with (80 – 20) Split**

Change the partition size to 80% train and 20% test.

A screenshot of a computer

Description automatically generated

**Interpretation**

Accuracy is 76%

Sensitivity is 57.50%

Specificity is 78.06%

Error Rate: 1- Accuracy which is 24%

**Fig24 Logistic\_ConfusionMatrix\_80-20\_spl 1**

**Cost Matrix**

* + **A close-up of a number

    Description automatically generated**The model incorrectly classified **186 loans as likely to default**, even though they wouldn’t. This misclassification could cause unnecessary concern for borrowers and might result in lost business opportunities for the bank.

**Fig25 Logistic\_CostMatrix\_80-20\_split 1**

* Additionally, it **missed 34 loans that did default**, leading to potential financial losses for the bank, as these high-risk customers were not identified for additional scrutiny or intervention.

**Decision Tree**

Here, we are using a decision tree with 60% Train and 40% Test data.

**A diagram of a computer network

Description automatically generated**

**Fig26 Decision Tree 1**

* High Debt-to-Income Ratio (>45) strongly predicts negative outcomes (90.5% probability).
* Low Delinquency (<=0.1) combined with higher Credit Lines (>2.746) increases the chance of a positive outcome to 77.55%.
* Delinquency (>0.1) is a critical factor in reducing the likelihood of a positive outcome, with an 80% chance of a negative result.
* The number of Derogatory Records and Credit Lines are significant secondary factors influencing the outcome.

**Evaluation Metrics**

**A screenshot of a computer

Description automatically generated**

**Interpretation**

Accuracy is 87%

Sensitivity is 48.15%

Specificity is 90.68%

Error Rate: 1- Accuracy which is 13%

**Fig27 DecisonTre\_ConfusnMatx\_60-40\_spt 1**

**Cost Matrix**

* **A close-up of a computer code

  Description automatically generated**The model incorrectly classified **158 loans as likely to default**, even though they wouldn’t. This misclassification could cause unnecessary concern for borrowers and might result in lost business opportunities for the bank.
* Additionally, it **missed 83 loans that did default**, leading to potential financial losses for the bank, as these high-risk customers were not identified for additional scrutiny or intervention.

**Fig28 DecisonTree\_CostMatrix\_60-40 1**

# COMPARISON OF MODEL PERFORMANCE

From the figure below, we can see that the decision tree with a split of 60% and 40% is performing well compared to other models in terms of accuracy, which is 87%. Although it has low sensitivity, it has the lowest cost function value out of all Models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **`** | **Accuracy** | **Sensitivity** | **Specificity** | **Error Rate** | **Missclassification** |
| **Logistic Regression (60-40)** | 78.93% | 62.50% | 80.48% | 22% | 1,13,161 |
| **Logistic Regression (70-30)** | 78.0% | 56.66% | 80.18% | 22% | 66,208 |
| **Logistic Regression (80-20)** | 76.0% | 57.50% | 78.06% | 24% | 35,752 |
| **Decision Tree (60-40)** | 87.0% | 48.15% | 90.68% | 13% | 31,853 |

**Fig29 Comparison of Model Performance 1**

# COST FUNCTION

A cost function measures the financial impact of misclassifications in a model by assigning specific costs to False Positives (FP) and False Negatives (FN). It helps evaluate how costly model errors are in real-world terms, guiding efforts to improve accuracy and minimize losses.

For the chosen **Decision Tree model using a 60-40** training and testing data split, we calculated the financial impact of the model's misclassifications. Here's a simple breakdown:

* FN = 158
* FP = 83
* C(FP) = $3,270,600
* C(FN) = $1,518,500

Total Cost = C(FN) x FN + C(FP) x FP

Total Cost = $1,518,500 \* 158 + $3,270,600 \* 83

= 239,923,000 + 271,459,800

**= 511,382,800**

* **False Negatives (FN)**: These are cases where loans predicted as "safe" were actually defaults. Each FN costs **$1,518,500**, and the model misclassified **158 cases**, adding up to **$239,923,000**.
* **False Positives (FP)**: These are cases where loans predicted as "default" were actually safe. Each FP costs **$3,270,600**, and the model misclassified **83 cases**, totaling **$271,459,800**.

**Total Cost**: Adding these together gives a total misclassification cost of **$511,382,800**.

# PROFIT ANALYSIS

Business Profit is related to how many loans we have sanctioned to the customers, but if the model fails to classify them correctly, the loss will be very high to the bank with respect to the loan amount.

Here is the figure below showing how much the loan amount bank has sanctioned to the customers regardless of defaulters and non-defaulters.

**A computer screen shot of a tree

Description automatically generated**

**Profit Calculation**

**1. True Positives (TP):**

- The decision tree correctly predicted good loans that were successfully repaid.

Sum of TP loans = $1,484,700

- This amount represents the revenue from good loans the bank successfully identified and approved.

**2. True Negatives (TN):**

- The decision tree correctly rejected bad loans that would have defaulted.

Sum of TN loans = $29,423,600

- This amount represents the money saved by the bank from not approving risky loans that would have defaulted.

**3. False Positives (FP):**

- The decision tree wrongly predicted bad loans as good, leading to defaults.

Sum of FP loans = $3,270,600

- This amount represents the financial loss from approving loans that defaulted.

**4. False Negatives (FN):**

- The decision tree wrongly rejected good loans, which resulted in missed opportunities.

Sum of FN loans = $1,518,500

- This amount represents the missed revenue from customers who would have repaid their loans but were incorrectly denied.

**Net Business Profit Calculation**

**Business Profit = Sum of TP Loans + Sum of TN Loans - Sum of FP Loans + Sum of FN Loans**

= $1,484,700 + $29,423,600 - ($3,270,600 + $1,518,500)

**= $26,119,200**

# BUSINESS INTERPRETATION

* The total business profit from using the decision tree model is **$26,119,200.**
* This profit comes from correctly identifying good loans (TP) and avoiding risky loans (TN), which resulted in significant revenue and cost savings.

**However, the decision tree model also caused:**

* A loss of **$3,270,600** from False Positives (approving loans that defaulted).
* A missed revenue of **$1,518,500** from False Negatives (rejecting loans that could have been repaid).

# OVERALL BUSINESS IMPACT

By following these recommendations, the bank can save more money by avoiding bad loans and making more money by approving good loans. This will help the bank reduce financial losses and increase profits—right now. The bank’s total profit from using this model is $26,119,200, but improving the model can increase that even further.

# RECOMMENDATIONS FOR IMPROVING LOAN DECISIONS

**1. Focus on avoiding bad loans (False Positives):**

The bank should focus on identifying customers who are likely to default to avoid financial losses. Approving bad loans can be very costly.

**Impact** In the current model, the bank lost $3,270,600 because it approved loans for customers who ended up not paying them back. Reducing these bad loans will help the bank save money.

**2. Maximize approval of good loans (False Negatives):**

Ensure that the bank doesn’t reject loans for customers who would have repaid, as this means missing out on potential profits.

**Impact** The bank missed out on $1,518,500 in revenue from customers who would have repaid their loans but were incorrectly denied. The bank can make more money by improving the model to catch more of these good loans.

**3. Balance risk and reward:**

The bank needs to find a good balance between avoiding bad loans and approving good loans. Too many bad loans lead to losses, but rejecting too many good loans means missing opportunities.

**Impact** While the model saved the bank $29,423,600 by correctly rejecting bad loans, it could have earned more by approving the good ones that were missed. Balancing this better will maximize profit.

**4. Give more importance to avoiding bad loans:**

It’s more important to avoid bad loans than it is to avoid missing a few good loans because the cost of a bad loan is much higher.

**Impact** The bank lost $3,270,600 from bad loans, which is more than the $1,518,500 it lost by missing good loans. The model should focus more on avoiding bad loans, as they are more harmful to the bank’s bottom line.

**5. Regularly update and improve the model:**

The bank should continuously improve the model to adapt to changing trends in customer behavior, ensuring it remains effective at predicting both good and bad loans.

**Impact** By regularly updating the model and testing it on new data, the bank can make sure it continues to reduce risk and increase profits over time.

# CONCLUSION

In this project, we analyzed loan data to understand the factors influencing loan defaults. By applying classification models, such as logistic regression and decision trees, we aimed to predict which customers are more likely to default on their loans. Through feature selection methods like Boruta, we identified key predictors, such as `CLAGE`, `Credit\_Lines`, and `LOAN`, which contribute to default risk. This analysis helps in identifying high-risk customers, allowing lenders to implement better risk management strategies and improve loan approval processes.

Decision Tree, with 60% to 40% portioned data has performed well compared to other logistic models with different split criteria. Although the accuracy is 87%, it has sensitivity of 48% but it has a low-cost function value which is 31,853, and low error rate. Since cost calculation is one of the important measures, we are considering that the decision tree has good results.

The key takeaway is to improve the model for better results otherwise the cost incurred for the loss will be very high for the bank.

# FUTURE SCOPE

For future work, we can explore advanced models like random forests or boosting algorithms to improve prediction accuracy. We could also try tuning the model parameters to enhance performance and further reduce errors. Additionally, incorporating more features or external data sources may help the model better understand customer behavior and provide even more reliable results.

**Name:** Priyanka Gonugunta

**Project:** Bank Loan Default Prediction

# Executive Summary

**Objective**

The project aims to identify key factors contributing to loan defaults and develop a predictive model to help the bank assess loan applicants' risk. This will improve decision-making, reduce default rates, and minimize financial losses for the bank.

**Approach**

We first cleaned the data by fixing any missing or incorrect information and then turned text-based data like job type and loan reason into a format the model could understand. Next, we explored the data to find patterns and relationships, especially focusing on how different factors like loan amount and credit history relate to defaults. After preparing the data, we used machine learning techniques to build a model that predicts whether someone is likely to default on a loan. Finally, test the model to make sure it accurately identifies people who are at high risk of defaulting.

**INTRODUCTION**

This project takes a close look at a dataset called "Consumer.csv" to understand how people shop and what influences their choices. The goal is to help companies create offers and promotions that bring customers back. In the past, companies focused on having a lot of products to sell, but now, the ones that stand out are those that truly understand their customers and use that knowledge to make better marketing decisions.

We start by analyzing the data with simple stats and visuals to spot patterns. Then, we divide customers into groups based on their shopping habits, like brand loyalty and purchase motivation. Next, we build models to identify "value-conscious" shoppers who hunt for deals and predict which customers might become loyal to specific brands. These insights help companies target the right offers to each group, whether by focusing on value or building brand loyalty.

By the end, the analysis will give companies a clear picture of their different types of customers and how to connect with them. Whether through discounts for deal-seekers or rewards for loyal customers, companies can create more meaningful promotions that keep customers coming back, helping them grow in the market.

**Business Problem**

The goal is to help AXANTEUS understand different types of customers based on their shopping habits and loyalty to brands. This way, AXANTEUS can give its clients, like advertising agencies and product manufacturers, useful insights on how to create better-targeted promotions. With this information, clients can reach the right groups of customers more effectively, making their advertising money work smarter and helping keep customers loyal.

**Cause**

Right now, AXANTEUS uses a basic approach to divide customers, mostly just by looking at general information like age or location. This doesn’t dig deep into how often people buy, which brands they stick with, or how much they care about discounts. Because of this, AXANTEUS misses important details that could help identify unique customer groups more accurately.

**Effect**

Since this approach isn’t very detailed, AXANTEUS’s clients might end up spending too much on broad promotions that don’t speak to any specific group. This can waste money, miss the chance to build loyal customer groups, and even allow competitors to gain an edge by targeting customers in a smarter, data-driven way.

**Business Goal**

The objective is to help AXANTEUS identify different groups of customers based on their shopping habits and brand loyalty. By understanding these groups, AXANTEUS can help its clients create more targeted promotions, boosting customer loyalty and making marketing spending more effective.

**Analytics Goal**

The main goal is to understand customer shopping patterns and determine what influences brand loyalty and responsiveness to promotions. Using a dataset of consumer purchases with 600 observations and 46 variables, we’ll look for key trends, identify the most impactful factors, and categorize customers into distinct groups.

Our analysis will involve three models. First, we’ll create a segmentation model that groups customers based on their shopping behaviors and purchase patterns (are they brand centric & price centric & both), giving AXANTEUS a clearer picture of customer types.

Next, we’ll build a value-consciousness model to identify customers who are highly responsive to deals and discounts based on the clustering results, helping AXANTEUS’s clients target cost-saving shoppers more effectively.

Finally, we’ll develop a brand loyalty prediction model that forecasts which customers are likely to repeatedly buy the same brand or not, identifying potential loyalists who can be nurtured for long-term retention.

With these models, AXANTEUS’s clients can make smarter, data-driven decisions in their marketing and promotions. By understanding which customers are deal-seekers, brand loyalists, or have unique shopping motivations, clients can craft targeted strategies to boost customer engagement, loyalty, and sales.

**Analytics Approach**

Our approach begins by exploring and understanding the consumer purchase data through visualizations and basic statistics to see patterns and trends in shopping behavior. Then, we’ll use a series of models to gain deeper insights: first, a segmentation model to categorize customers into different groups based on their shopping habits and motivations, helping us understand unique customer types.

After that, we’ll develop a value-consciousness model to identify customers who are especially responsive to deals and discounts. Finally, we’ll build a brand loyalty prediction model to forecast which customers are likely to stick with certain brands over time.

Together, these models will give AXANTEUS’s clients a data-driven foundation for making smarter marketing decisions, ensuring their promotions and offers are more effectively targeted to the right customer groups.

DATA PRE-PROCESSING

**VARIABLE DESCRIPTION**

**Member ID:** Unique identifier for each consumer.

**SEC (Socioeconomic Class**): Indicates the socioeconomic status of the consumer (1 = High, 5 = Low).

**FEH (Eating Habit):** Describes dietary habits (1 = Vegetarian, 2 = Vegetarian but eats eggs, 3 = non-vegetarian, 0 = Not Specified).

**MT (Native Language**): Language code representing the consumer’s native language (0-19).

**SEX:** Gender of the consumer (1 = Male, 2 = Female).

**AGE:** Age group of the consumer.

**EDU (Education Level):** Education level of the consumer (1 = Minimum, 9 = Maximum).

**HS (Household Size**): Number of people in the consumer’s household.

**CHILD:** Number of children in the consumer’s household.

**CS (Television Availability):** Indicates whether the consumer has a television (1 = Available, 2 = Not Available).

**Affluence Index:** Weighted score based on the consumer's possession of durable goods, reflecting their economic status.

**No. of Brands:** Number of different brands purchased by the consumer.

**Brand Runs:** Frequency of consecutive purchases for the same brand, indicating brand loyalty.

**Total Volume:** Total amount of products purchased by the consumer.

**No. of Trans (Number of Transactions):** Total purchase transactions made by the consumer over the period.

**Value:** Total spending amount by the consumer.

**Trans / Brand Runs:** Average number of transactions per brand run, indicating purchase consistency.

**Vol/Tran (Volume per Transaction):** Average amount of product purchased in each transaction.

**Avg. Price:** Average price paid per purchase by the consumer.

**Promotion Variables:**

- Pur Vol No Promo %: Percentage of volume purchased without any promotion.

- Pur Vol Promo 6 %: Percentage of volume purchased under promotion code 6.

- Pur Vol Other Promo %: Percentage of volume purchased under other promotions.

**Brand-Specific Purchases:**

- Variables like Br. Cd. 57, Br. Cd. 55, Br. Cd. 272, etc., represent the percentage of volume purchased under specific brand codes.

**Price Category Variables:**

- Pr Cat 1, Pr Cat 2, etc.: Percentage of volume purchased under specific price categories (1–4), reflecting pricing preferences.

**Selling Proposition Variables:**

- PropCat 5, PropCat 6, etc.: Percentage of volume purchased under specific selling proposition categories (5–15), indicating consumer preference based on the selling proposition.

**Renaming the Variables**

In this section, we are renaming the following variables to make them more understandable and meaningful.

**Demographic Variables**

* Member.id has been renamed to MemberID.
* SEC has been renamed to SocioEconomic.
* FEH has been renamed FoodHabit.
* MT has been renamed to NativeLanguage.
* EDU has been renamed to Education.
* HS has been renamed to FamilyMembers.
* CHILD has been renamed to Child.
* CS has been renamed TelevisionAvailability.
* Affluence Index has been renamed to AffluenceIndex.
* No. of Brands has been renamed to NumOfBrands.
* Brand Runs has been renamed to BrandRuns.
* Total Volume has been renamed TotalVolume
* No. of Trans has been renamed to Invoices.
* Value has been renamed to Revenue.
* Trans / Brand Runs has been renamed to trans\_per\_brand\_run.
* Vol/Tran has been renamed to volPerTran.
* Avg. Price has been renamed to AvgPrice.

**Promotion Variables**

* Pur Vol No Promo - % has been renamed to pur\_vol\_no\_promo\_pct
* Pur Vol Promo 6 % has been renamed to pur\_vol\_promo\_6\_pct
* Pur Vol Other Promo % has been renamed to pur\_vol\_other\_promo\_pct

**Brand-Specific Purchases**

* Br. Cd. 57, 144 has been renamed to brand\_cd\_57\_144\_pct
* Br. Cd. 55 has been renamed to brand\_cd\_55\_pct
* Br. Cd. 272 has been renamed to brand\_cd\_272\_pct
* Br. Cd. 286 has been renamed to brand\_cd\_286\_pct
* Br. Cd. 24 has been renamed to brand\_cd\_24\_pct
* Br. Cd. 481 has been renamed to brand\_cd\_481\_pct
* Br. Cd. 352 has been renamed to brand\_cd\_352\_pct
* Br. Cd. 5 has been renamed to brand\_cd\_5\_pct
* Others 999 has been renamed to others\_999\_pct to represent other brands

**Price Category Variables**

* Pr Cat 1 has been renamed to price\_cat\_1\_pct
* Pr Cat 2 has been renamed to price\_cat\_2\_pct
* Pr Cat 3 has been renamed to price\_cat\_3\_pct
* Pr Cat 4 has been renamed to price\_cat\_4\_pct

**Selling Proposition Variables**

* PropCat 5 has been renamed to prop\_cat\_5\_pct
* PropCat 6 has been renamed to prop\_cat\_6\_pct
* PropCat 7 has been renamed to prop\_cat\_7\_pct
* PropCat 8 has been renamed to prop\_cat\_8\_pct
* PropCat 9 has been renamed to prop\_cat\_9\_pct
* PropCat 10 has been renamed to prop\_cat\_10\_pct
* PropCat 11 has been renamed to prop\_cat\_11\_pct
* PropCat 12 has been renamed to prop\_cat\_12\_pct
* PropCat 13 has been renamed to prop\_cat\_13\_pct
* PropCat 14 has been renamed to prop\_cat\_14\_pct
* PropCat 15 has been renamed to prop\_cat\_15\_pct

**Structure of the Data**

We have a total of 600 observations and 46 variables in our dataset, out of 46 variables, 16 are integer variables, and 30 are numerical variables.

**A screenshot of a computer program

Description automatically generated**  **A screenshot of a computer code

Description automatically generated**

**Checking for the Missing Values and Null Values**

There are no Missing Values and Null Values in the data.

A close-up of a white background

Description automatically generated

**Checking for the Zeros in the column**

Although there are some zeros in the columns, we are taking these variables without handling because they are giving meaningful information.

**Summary of Overall Data**

**A close up of a text

Description automatically generated**

UNI – VARIATE ANALYSIS

**Demographic Variables**

1. **A graph of a number of bars

   Description automatically generatedSocioeconomic Class**

Where,

* 1 = High
* 4 = Low

From the graph, we can see that the data contains a balanced number of consumers from each socioeconomic class.

Fig1 Socioeconomic Class 1

1. A graph with blue rectangles

   Description automatically generated**Eating Habits (FEH)**

Where,

* 0: Not Specified
* 1: Vegetarian
* 2: Vegetarian but eats eggs
* 3: Non-vegetarian

Fig2 Eating Habits 1

This chart shows that most people in the dataset are non-vegetarians, with a smaller number being vegetarians.

1. A graph of a number of different languages

   Description automatically generated**Native Language**

where,

0~19 language code

Fig3 Native Language 1

From the above figure, we can say that most people in the dataset (over 300) speak the language represented by code **10**.

* A few other languages, such as those represented by codes **0**, **3**, and **6**, have some speakers, but in much smaller numbers.
* Many of the language codes have very few or no speakers.

1. **Gender of Consumer**

A graph with a bar and a person in the background

Description automatically generated with medium confidence

where,

* 1 = Male
* 2 = Female
* 0 = Unspecified

Fig4 Gender 1

This distribution suggests that the majority of consumers in this dataset are female. This information can help AXANTEUS, and its clients focus their marketing strategies to appeal more to female consumers, as they make up the largest group.

1. **Age**

A graph of a number of people

Description automatically generated

* The largest group of consumers falls into the underage category **4**, with close to 300 people.
* Age categories **2** and **3** also have a significant number of people, with around 150–200 consumers each.
* Very few consumers are in category **1**.

Fig5 Age 1

1. **Education**

**A graph of a graph of a number of people

Description automatically generated with medium confidence**

Where,

Consumer Education (1 = Minimum, 9 = Maximum)

* Most consumers fall into education level **5**, with over 150 people in this group.
* Education levels **4** and **6** are also common, with around 100–150 people each.
* Fewer consumers have very low (0, 1, 2) or very high (8, 9) education levels.

Fig6 Education 1

1. **Number of members of Consumer household**

**A graph of a family member

Description automatically generated**

* Most consumers have a household size of 4 or 5 members, with each of these categories having around 150 people.
* There are fewer households with either very small or very large sizes.
* Household sizes mostly range from 1 to 7 members, with a small number going beyond that.

Fig7 No.of Members in HouseHold 1 1

1. **CHILD**

A graph of a number of children

Description automatically generated

* Most households have **4 children**, with over 250 in this category, making it the most common number of children.
* Households with **2 children** are the second most common, with around 150.
* Fewer households have 1, 3, or 5 children.

Fig8 Child 1

1. **Television Availability**

A graph with a bar

Description automatically generated

* The majority of households (over 400) have a television, represented by **1**.
* Around 100 households do not have a television, represented by **0**.
* A small number of households fall into a separate category, represented by **2** (which may indicate partial or other types of television availability).

Fig9 Television Availability 1

1. **A graph with blue bars

   Description automatically generatedAffluence Index**

* Most consumers have an Affluence Index between 0 and 30, with the highest count around the 10 to 15 range. This means a majority of consumers have a lower to mid-level affluence score.
* As the Affluence Index increases (above 30), the number of consumers drops, meaning fewer consumers have high levels of wealth.

Fig10 Affluence Index 1

**Purchase Summary Over the Period**

1. A green bar graph with black line

   Description automatically generated**Number of Brand Purchased**

* Most consumers buy between **2 to 4 different brands**, with the largest group buying **3 brands**.
* Fewer consumers buy a wide variety of brands, with very few purchasing **8 or 9 brands**.
* The count decreases as the number of brands purchased increases, meaning that only a small number of consumers buy from many different brands.

Fig11 No of Brand Purchased 1

1. **A graph of a graph

   Description automatically generated with medium confidenceBrand Runs**

* Most consumers stick with the same brand for around 4 to 18 consecutive purchases, with peaks at 4, 10, and 16.
* Beyond 20 purchases, fewer consumers remain loyal to the same brand. Very long brand runs (30+ consecutive purchases) are rare.

Fig12 Brand Runs 1

1. **Total Volume Purchased**

A purple graph with black lines

Description automatically generated

* Most consumers purchased a total volume of between **5,000 and 15,000 units**.
* The highest number of consumers is around **10,000 units**.
* Only a few consumers bought very large volumes (above **40,000 units**), making high-volume buyers rare.

Fig13 Total Volume Purchased 1

1. **Value**

**A graph of a number of bars

Description automatically generated**

* Most consumers have a purchase value between 500 and 2,000 units, with peaks around 1,000.
* Only a few consumers have a high purchase value above 4,000.
* The spending rate generally decreases as the value gets higher, meaning fewer people make very large purchases.

Fig14 Value 1

1. A graph with a blue line

   Description automatically generated**Average Transaction per Brand Run**

Fg15 Average Transaction per Brand Run 1

Most consumers have a very low number of transactions per brand run, shown by the thick line in the middle. However, a few consumers make a lot of consecutive purchases from the same brand, which are shown as dots above. Few consumers have purchased more than the typical range(Mean).

1. A graph with a line

   Description automatically generated**Volume Per Transaction**

Most people buy around 500 items per transaction, but some buy a lot more in one go. These unusually large purchases which are indicated as dots.

Fig16 Volume Per Transaction 1

1. A graph with a line and dots

   Description automatically generated**Average of Purchase**

Most consumers spend around 10 units per item on average, but some pay much higher prices which are indicated as dots.

Fig17 Average of Purchase 1

1. **A graph of a number of invoices

   Description automatically generatedNumber of Transactions**

* Most people in this dataset have made around 25 transactions, with only a few making many purchases.
* The frequency decreases as the number of transactions increases, meaning fewer people make a very high number of transactions. Only a small number of consumers have made over 50 transactions.

Fig18 Number of Transactions 1

**Purchase Within Promotion**

1. **Percent of Value with No Promotion**

**A graph of a graph

Description automatically generated with medium confidence**

From the above figure, we can say that most of the spending happens without promotions. Promotion Code 6 and other promotions account for only a small percentage of total spending for most consumers.

Fig19 Percent of Value with No Promotion 1

1. **Brand Wise Purchase**

**A group of graphs showing the value of a number of different colored lines

Description automatically generated with medium confidence**

*Fig20 BrandWise Purchase 1*

* **Brand Codes 57, 55, 272, 286, 24, 481, 352, and 5:** Most consumers spend a very small percentage on each of these specific brands, which is almost near to 0%
* **Others (999):** Spending in the "Others" category varies widely, which means consumers are spending money across many smaller brands in the "Others" category.

1. **Price Category with Purchase**

**A group of graphs with numbers and lines

Description automatically generated with medium confidence**

* Pr Cat 3 and Pr Cat 4: Most consumers spend very little on these categories.
* Pr Cat 1: Consumers spend a mix of amounts here but generally tend to make smaller purchases.
* Pr Cat 2: Spending is more spread out, with consumers spending a wide range of percentages in this category, from small to large amounts.

Fig21 Price Category with Purchase 1

1. **Selling Proportion wise Purchase**

**A graph of a graph

Description automatically generated with medium confidence**

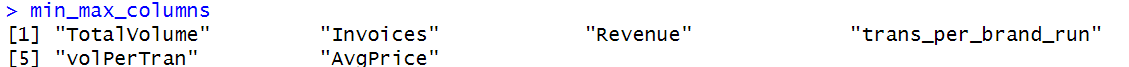
*Fig22 Selling Proportion wise Purchase 1*

* Consumers tend to spend a larger portion on PropCat 5 compared to other categories. For the remaining categories from PropCat 6 to PropCat 15, spending is minimal, with most consumers allocating only a small percentage of their purchases to these categories.

DATA TRANSFORMATION

**Scaling the Data**

In our project, scaling ensures that different types of consumer data, such as "Total Volume" and "Average Price," are on a comparable range before clustering. Without scaling, larger numbers (like total volume) might dominate the analysis, making it difficult to accurately capture patterns across all variables. By scaling each variable to a similar range, we ensure that every piece of information whether it's about spending, brand loyalty, or socio-economic factors contributes fairly to the clustering results. This step is essential to make meaningful groupings based on consumer behaviors and characteristics.



We are using Z-score scaling to transform our data values to a range between -1 and 1. This approach is helpful because it standardizes the scale of all variables, allowing each feature to contribute proportionately to our clustering analysis, regardless of its original range.

DIMENSION REDUCTION

In this project, although PCA could help reduce the dataset's dimensionality by highlighting the main factors affecting consumer behavior, we have decided not to implement it. Our focus is on maintaining the interpretability of the original variables, as PCA can obscure which specific variables drive cluster formation. By using all relevant variables directly, we preserve clarity in our analysis and ensure we understand the role of each variable in consumer segmentation. This approach avoids the complexity of mapping back from principal components to the original data, making our clustering results easier to interpret.

DATA PARTITIONING

When working with a dataset of only 600 observations, partitioning the data (such as dividing it into training and testing sets) generally does not bring notable benefits for clustering stability or robustness. Clustering models, especially in unsupervised learning tasks, relieve identifying patterns and relationships within the entire dataset, and they are often quite stable even with smaller datasets. With only 600 observations, there is usually enough information within the whole dataset to create meaningful clusters without the need to partition it.

Partitioning is primarily beneficial when we have a large amount of data, as it helps to validate model performance across different subsets. In smaller datasets, like ours, removing part of the data for validation can sometimes weaken the clustering results, as each data point contributes valuable information about the underlying patterns. Therefore, in this case, using the full dataset for clustering allows for the most robust and accurate identification of natural groupings within the data.

But we are using partitioning methods for our classification and regression models.

**Partitioning for classification Model**

Here we are taking 70% as train and 30% test data and value conscious as a target variable

**A computer screen shot of a computer code

Description automatically generated**

**Partitioning for Regression Model**

Here we are taking 80% as train and 20% test data and average price as a target variable

**A computer screen shot of a program

Description automatically generated**

MODEL SELECTION

Since our main goal is to segment the consumers based on their purchase consciousness and behavior, here we are going to use an unsupervised technique called **Clustering**

After obtaining the clustering results, we will build two additional models to further analyze consumer behavior. These models will help us determine whether a consumer is value-conscious or brand-loyal. To achieve this, we plan to use a rule-based approach and a supervised model, specifically by implementing a **Decision Tree.** This approach allows us to classify consumers based on patterns identified in the clustering, providing actionable insights into consumer preferences and loyalty.

UN-SUPERVISED LEARNING

**PURCHASE BEHAVIOUR**

We chose these specific variables because they give a clear picture of consumer loyalty and spending habits.

**Key Variables:**

- No. of Brands and Brand Runs: Capture general brand loyalty patterns.

- Total Volume, No. of Trans, and Value: Represent overall purchasing levels.

- Trans / Brand Runs and Vol / Tran: Reflect buying frequency and purchase size.

- Avg. Price and Promo Percentages: Highlight price sensitivity and responsiveness to promotions.

A diagram of a cluster plot

Description automatically generated A graph with different colored bars

Description automatically generated with medium confidence

*Fig23 Purchase Behaviour Clustering 1*

**Interpretation:**

**Cluster 1**:

* These consumers stick to fewer brands and shop less frequently.
* They are less interested in promotions and the quantity of purchases for each transaction is high

**Cluster 2**:

* These consumers buy from more brands and shop more often for the same brands
* They prefer deals and promotions and bulk volume per transaction, making them value-conscious shoppers.

**Evaluation with Silhouette**

A screenshot of a computer

Description automatically generated The clustering results indicate a moderate level of separation between the groups. Cluster 1 is more cohesive, with a higher average silhouette score, while Cluster 2 is less well-defined and shows overlap. To enhance these results, we could experiment with a different number of clusters, refine the variables, or explore alternative clustering algorithms.

Fig24 Purchase\_Behaviour\_Silhouette 1

**Conclusion:**

Both clusters demonstrate brand loyalty with minimal promotion influence, but Cluster 2 shows a broader brand engagement and higher purchase volume than Cluster 1.

**BASIS FOR PURCHASE**

We chose these variables because they help us understand why consumers buy certain products:

1. **TotalVolume**: The total amount of purchases made by each consumer.
2. **trans\_per\_brand\_run**: The average number of transactions for each brand run.
3. **AvgPrice**: The average price paid by the consumer.
4. **prop\_cat\_5\_pct**: The proportion of purchases made in category 5.
5. **prop\_cat\_15\_pct**: The proportion of purchases made in category 15.

**A graph showing a plot of a product

Description automatically generated with medium confidence**These variables give us insight into what drives buying choices based on price and product type. **A screenshot of a computer screen

Description automatically generated**

*Fig25 Basis of Purchase Clustering 1*

**Interpretation**

**Cluster 1:** Consumers in this cluster show higher values in prop\_cat\_5\_pct and Average Price, indicating that they prefer products within category 5 and are willing to spend more on average. This suggests a focus on specific product categories with higher pricing.

**Cluster 2:** Consumers in this cluster are purchasing low value but more quantity and trans\_per\_brand\_run, indicating more transactions per brand run and a higher proportion of purchases in category 15. This suggests a preference for frequent purchases in a specific category.

**Evaluation with Silhouette**

A graph of a graph

Description automatically generated with medium confidence This silhouette plot shows two clusters with moderate separation. Cluster 1 has a silhouette score of 0.29, indicating somewhat clear grouping, while Cluster 2 has a lower score of 0.09, suggesting significant overlap with Cluster 1. Overall, the clustering model provides a basic separation but could be improved to achieve better-defined clusters.

Fig26 Basis\_of\_Purchase\_Silhouette 1 1

**Conclusion:**Cluster 1 represents consumers who prefer higher-priced products within specific categories, while Cluster 2 reflects consumers buying in bulk which has less price.

**PURCHASE BEHAVIOU AND BASIS OF PURCHASE**

1. **Basis of Purchase:** Average Price, Total\_Volume, Average Transaction Per Brand run and Product Propositions (`PropCat 5 & 15`) show consumer spending preferences and interests.

2. **Purchase Behavior:** Metrics like `Total Volume`, `No. of Trans`, `Value`, ‘Brand runs’,’No.of brands, ‘invoices reflect actual buying actions, loyalty, and frequency, helping us understand consumer habits and brand engagement.

Together, they reveal why consumers buy and how they engage with brands.

**A graph of a diagram

Description automatically generated with medium confidence**  A colorful rectangular shapes with text

Description automatically generated with medium confidence

*Fig27 Basis of Purchase & Behaviour Clus 1*

**Interpretation**

**Cluster 1**

* **Total Volume and Invoices:** Moderate values reflect moderate purchase activity.
* **Promotions:** Lower reliance on promotions (pur\_vol\_no\_promo\_pct is higher), showing less sensitivity to discounts or deals.
* **Brand and Product Propositions:** A more distributed preference across product categories (prop\_cat\_15\_pct is notable), indicating varied purchasing behavior.

**Cluster 2**

* **Total Volume and Invoices:** Higher values indicate consumers in this cluster purchase more frequently and in larger quantities.
* **Promotions:** Strong reliance on promotions (pur\_vol\_promo\_6\_pct and pur\_vol\_other\_promo\_pct are higher), suggesting value-conscious behavior.
* **Price and Product Propositions:** High proportions in certain product (prop\_cat\_5\_pct) and price categories suggest focused preferences in these areas.

**Evaluation with Silhouette**

A graph of a graph

Description automatically generated with medium confidence

This silhouette plot shows an average silhouette width of 0.3, which is relatively low, indicating that the clustering might not be well-separated. Cluster 1 has a slightly better silhouette score (0.42) compared to Cluster 2 (0.28), suggesting that Cluster 1's members are more tightly grouped, while Cluster 2 has more overlap or ambiguity with other clusters.

Fig28 Basis\_of\_Purchase&Behaviou\_Silhout 1

**Conclusion:**

Cluster 2 represents value-conscious consumers with high promotional reliance and focused product preferences.

Cluster 2 represents consumers with stable purchasing patterns and less focus on promotions.

PREDICTOR RELEVANCY

After obtaining insights from clustering, the **Value Consciousness Model** will predict customers' value orientation using demographic variables (e.g., age, income, location), excluding MemberID. This model helps identify value-focused customer segments and understand the factors influencing their purchasing decisions. Building on this, the **Potential Loyal Customers Model** classifies customers likely to demonstrate loyalty based on demographics and clustering insights. By leveraging these models, the business can enhance its targeting strategies, design personalized marketing campaigns, and foster customer retention, ultimately driving long-term growth and loyalty.

SUPERVISED LEARNING

**Classification**

Based on the results we got from 3 clustering models; we can derive value consciousness by focusing on clusters that exhibit high sensitivity to promotions and certain price or brand categories by using the results from 3rd clustering model (Basis of Purchase & Purchase Behaviour)

Based on these results labeled the data whether they are value conscious or not.

**A close-up of a sign

Description automatically generatedCreate the Labels**

From the figure, we can say that almost 41% which 248 out of 600 consumers are value conscious, which means they are checking for the deal or discounts and buying in large quantities.

**Logistic Regression on Value Conscious**

The value-conscious model created using logistic regression identifies consumers who prioritize promotions and discounts in their purchasing behavior. The model evaluates demographic variables such as socioeconomic attributes, gender, age, eating habits, television availability, number of family members etc., and promotional reliance to classify consumers as value-conscious or not. This helps in segmenting and targeting customers who are likely to respond to promotional strategies.

**A computer screen shot of blue text

Description automatically generated**

**A screenshot of a computer

Description automatically generatedEvaluation Metrics**

The logistic regression model has an accuracy of 65%, indicating that 65% of the predictions were correct. The sensitivity (0.4474) shows the model's ability to correctly identify value-conscious consumers (true positives), which is relatively low. The specificity (0.7981) measures the ability to correctly identify non-value-conscious consumers (true negatives), which is higher. The error rate, which is 1 minus accuracy, indicates that 35% of the predictions were incorrect. These metrics suggest that while the model performs well in identifying non-value-conscious consumers, it struggles with identifying value-conscious ones.

**Linear Regression on Average Price**

The Linear Regression model predicts the Average Price consumers pay based on Demographic variables. It identifies which variables (like purchase volume or product preferences) influence the price the most. This helps us understand spending patterns and target strategies for pricing or promotions accordingly.

**Blue text on a white background

Description automatically generated**

**Accuracy**

The R-squared value of **0.38** means that the model explains **38%** of the variation in average price based on the given variables. This indicates a moderate fit, suggesting the model captures some trends but could be improved to explain more variability.

**A close-up of a computer code

Description automatically generated**

**Evaluation Metrics**

**A white background with black text

Description automatically generated**The Root Mean Squared Error (RMSE) of 3.064003 means that, on average, the predictions made by the model deviate from the actual values by about 3 units. This value helps measure the accuracy of the model, and a lower RMSE indicates better performance. It shows the typical error magnitude in the model's predictions.

CONCLUSION

In this project, we applied clustering and regression techniques to uncover key consumer patterns and preferences. Using scaling methods like z-score standardization, we refined the data for effective analysis, ensuring that all variables contributed equally to the models. Our analysis revealed distinct consumer segments, including value-conscious and brand-loyal consumers, based on behavioral and purchase attributes.

Additionally, the logistic regression model helped classify value-conscious customers, while the linear regression model provided insights into variables affecting average prices. These findings offer actionable insights to support data-driven strategies for targeted marketing and consumer engagement.

RECOMMENDATIONS

* **BASED ON CLASSIFICATION**

**Enhance Sensitivity:**

The model's sensitivity is low (0.4474), indicating difficulty in identifying value-conscious consumers. Consider incorporating additional variables (e.g., past purchase frequency, online browsing behavior, or payment methods) to improve true positive identification**.**

**Leverage Specificity**:

The model performs well in identifying non-value-conscious consumers (specificity: 0.7981). Use this strength to exclude these consumers from heavy promotion campaigns, optimizing marketing spending.

* **BASED ON REGRESSION**

Use insights from the linear regression model to design pricing and promotional strategies that align with consumer spending habits. For example, they offer tiered pricing or volume-based discounts for consumers with higher purchase volumes.

Name: Priyanka Gonugunta

Project: Consumer Segmentation Analytics

Executive Summary

**Objective**

The goal of this project is to help AXANTEUS, a market research company, understand how different groups of people buy products and respond to discounts. By examining patterns in how often people buy certain brands, their loyalty to those brands, and how sensitive they are to promotions, we aim to divide consumers into different groups. This allows AXANTEUS’s clients to create more focused advertising and promotions, keeping customers coming back and boosting their share in the market.

**Approach**

We analyze data to find important patterns in consumer buying behavior, focusing on things like loyalty, spending habits, and interest in discounts. Using techniques like grouping similar consumers and predicting buying behaviors, we can identify trends and patterns. To make these patterns easy to understand, we use charts and diagrams that show the relationships in the data. These findings will help AXANTEUS’s clients target their promotions more accurately, ensuring that each customer group receives relevant offers and rewards, which can increase sales and build brand loyalty.

With these segments in place, we then build two predictive models to gain deeper insights. The first is a value-conscious model, which identifies consumers who are highly focused on deals and saving money, often by buying in bulk. This helps AXANTEUS’s clients target these consumers with promotions tailored to their cost-saving preferences. The second model is a brand loyalty prediction model that forecasts "Brand Runs"—patterns where consumers consistently purchase the same brand. This model helps clients spot potential loyal customers, allowing them to engage these consumers more effectively and encourage long-term loyalty.