**George Herbert Walker School of Business & Technology**

# Department of Computer & Information Sciences

# CSDA 6010 Data Analytics Practicum



# Case Studies:

* Maximizing Profits from Software Sales
* Empowering Used Mobile-Device Buyer
* Mortgage Payback Behavior Analytics

# 

**May 2024**

**Saint Louis, Missouri**

Department Chair: Dr. James Curtis

Instructors:

Dr. Jiangping Wang   
Dr. Ali Ovlia

CSDA 6010 Data Analytics Practicum

Sreeja Reddy Singidi

Webster University

Spring 2024

# **PROJECT 1**: **MAXIMIZING PROFITS FROM SOFTWARE SALES**

**TABLE OF CONTENTS**

[1. Introduction 13](#_Toc165902064)

[2. Project Overview 13](#_Toc165902065)

[3. Business Understanding 14](#_Toc165902066)

[4. Business Goals 14](#_Toc165902067)

[5. Attributes Definition 14](#_Toc165902068)

[6. Data Understanding 16](#_Toc165902069)

[6.1 Structure 16](#_Toc165902070)

[6.2 Dimension 16](#_Toc165902071)

[6.3 Checking For Missing Values 16](#_Toc165902072)

[6.4 Checking For Zero Values 16](#_Toc165902073)

[6.5 Attribute Analysis 16](#_Toc165902074)

[6.5.1 Distribution Of Outcome Variables 17](#_Toc165902075)

[6.5.2 Distribution Of Categorical Variables 17](#_Toc165902076)

[6.5.3 Distribution Of Numeric Attributes 19](#_Toc165902077)

[6.5.4 Distribution Of All Sources When Purchased 19](#_Toc165902078)

[6.5.5 Analyzing The Variable Frequency With Outcome Variables 20](#_Toc165902079)

[6.5.6 Average Spending By Gender 21](#_Toc165902080)

[6.5.7 Boxplots For Numeric Attribute 21](#_Toc165902081)

[6.5.8 Scatterplot Matrix To Know The Relation Between Each Numeric Attribute 21](#_Toc165902082)

[7. Predictor Analysis And Relevancy 22](#_Toc165902083)

[8. Data Transformations 22](#_Toc165902084)

[9. Dimension Reduction 23](#_Toc165902085)

[10. Business Considerations 23](#_Toc165902086)

[11. Data Partitioning 24](#_Toc165902087)

[12. Models Selection 24](#_Toc165902091)

[12.1 Goal 1- Classification 24](#_Toc165902092)

[12.1.1 Logistic Regression 25](#_Toc165902093)

[12.1.2 Backward Method 25](#_Toc165902094)

[12.1.3 Classification Tree 26](#_Toc165902095)

[12.1.4 K-Nearest Neighbor 28](#_Toc165902096)

[Compare The Classification Models With Accuracy And Sensitivity 28](#_Toc165902097)

[12.2 Goal 2- Regression 28](#_Toc165902098)

[12.2.1 Linear With Stepwise Regression 29](#_Toc165902099)

[12.2.2 Regression Trees 29](#_Toc165902100)

[Comparing Models For Predicting Spending 30](#_Toc165902101)

[Models Selected To Predict On Holdout Data 30](#_Toc165902102)

[13. Holdout Data Analysis And Adjustments: 30](#_Toc165902103)

[13.1 Comparing Logistic Regression On Validation And Test Data 30](#_Toc165902104)

[13.2 Comparing Linear Regression On Validation And Test Data 30](#_Toc165902105)

[14 Profit Analysis 31](#_Toc165902106)

[14.1 Gross Profit 31](#_Toc165902107)

[14.2 Predicted Probability Of Purchase 31](#_Toc165902108)

[14.3 Adjusted Probability Of Purchase 31](#_Toc165902109)

[14.4 Predicted Spending Value 31](#_Toc165902110)

[14.5 Expected Spending 32](#_Toc165902111)

[14.6 Cumulative Gain Chart 32](#_Toc165902112)

[15 Conclusion 32](#_Toc165902113)

[16 Business Recommendations 32](#_Toc165902114)

[Executive Summary 33](#_Toc165902115)

# **LIST OF FIGURES**

Figure 1: Overview. 13

Figure 2: Distribution of Outcome Variable 17

Figure 3: Distribution of categorical variables. 18

Figure 4: Distribution of numeric attributes. 19

Figure 5: Distribution of numeric attributes. 19

Figure 6: Distribution of Purchase with Frequency 20

Figure 7: Distribution of Spending with Frequency 20

Figure 8: Average Spending by Gender 21

Figure 9: Boxplots for numeric attribute. 21

Figure 10: Scatterplot Matrix 22

Figure 11: Classification tree with all predictors. 27

Figure 12: Classification tree with important predictors. 27

Figure 13: Regression Tree 29

Figure 14: Cumulative gain chart. 32

# **LIST OF TABLES**

[Table 1: Attributes 15](#_Toc165794164)

[Table 2: New Column Names 22](#_Toc165794165)

[Table 3: Logistic Regression Models 26](#_Toc165794166)

[Table 4: Classification model selection 28](#_Toc165794167)

[Table 5: Comparing models for predicting spending. 30](#_Toc165794168)

[Table 6: Comparing logistic regression on validation and test data. 30](#_Toc165794169)

[Table 7: Comparing linear regression with backward on validation and test data. 30](#_Toc165794170)

# **PROJECT 2: Empowering Used Mobile-Device Buyer**

**TABLE OF CONTENTS**

[1. Introduction 35](#_Toc166076926)

[2. Project Overview 35](#_Toc166076927)

[3. Business Understanding 35](#_Toc166076928)

[4. Business Goal 36](#_Toc166076929)

[5. Analytical Approach 36](#_Toc166076930)

[6. Attribute Definition 37](#_Toc166076931)

[7. Data Understanding 39](#_Toc166076932)

[7.1 Summary Statistics 39](#_Toc166076933)

[7.2 Checking For Missing Values 40](#_Toc166076934)

[7.3 Checking For 0 Values 41](#_Toc166076935)

[7.4 Attribute Analysis 41](#_Toc166076936)

[7.4.1 Distribution Of Target Variable 41](#_Toc166076937)

[7.4.2 Categorical Attributes 42](#_Toc166076938)

[7.4.3 Numerical Attributes 44](#_Toc166076939)

[8 Predictor Analysis And Relevancy 49](#_Toc166076940)

[8.1 Analysis 49](#_Toc166076941)

[8.1.1 Average Price By Ram 49](#_Toc166076942)

[8.1.2 Relation Between The Used Price And A New Price 49](#_Toc166076943)

[8.1.3 Correlation Between The Numeric Data 50](#_Toc166076944)

[8.1.4 Price Per Brand 51](#_Toc166076945)

[8.1.5 Price Per Os 51](#_Toc166076946)

[8.2 Relevancy 52](#_Toc166076947)

[8.2.1 Lasso Model 52](#_Toc166076948)

[8.2.2 Boruta Method 52](#_Toc166076949)

[8.2.3 Stepwise Regression 53](#_Toc166076950)

[9 Data Transformation 54](#_Toc166076951)

[9.1 Changing Column Name 54](#_Toc166076952)

[10 Dimension Reduction 54](#_Toc166076953)

[10.1 Combining Network Type 54](#_Toc166076954)

[11 Data Partitioning Methods 54](#_Toc166076955)

[Necessity 54](#_Toc166076956)

[Different Approaches 54](#_Toc166076957)

[Approach In This Project 54](#_Toc166076958)

[Importance 55](#_Toc166076959)

[12 Tablet And Phone Datasets 55](#_Toc166076960)

[13 Goal 1: Predicting The Price Of A Used Device 55](#_Toc166076961)

[13.1 Data Preparation For Regression Models 55](#_Toc166076962)

[Step 1: 55](#_Toc166076963)

[Step 2: Dividing The Dataset Into Phone And Tablets 57](#_Toc166076964)

[Step 3: Partitioning The Datasets. 58](#_Toc166076965)

[13.2 Regression Models 58](#_Toc166076966)

[13.2.1 Linear Regression Model 58](#_Toc166076967)

[13.2.2 Stepwise Regression 58](#_Toc166076968)

[13.2.3 Regression Tree 58](#_Toc166076969)

[13.3 Regression Model Selection 60](#_Toc166076970)

[14 Goal 2: Classifying Used Device Price Categories 60](#_Toc166076971)

[14.1 Data Preparation For Classification Models 61](#_Toc166076972)

[Step 1: 61](#_Toc166076973)

[Step 2: Removing Unnecessary Columns 61](#_Toc166076974)

[Step 3: Tablet And Phone Division 61](#_Toc166076975)

[Step 4: Partitioning The Datasets. 61](#_Toc166076976)

[14.2 Classification Models 62](#_Toc166076977)

[14.2.1 Logistic Regression 62](#_Toc166076978)

[14.2.2 Decision Tree 62](#_Toc166076979)

[14.2.3 K-Nearest Neighbors 62](#_Toc166076980)

[15 New And Used Price Comparison For Phone And Tablet Datasets 63](#_Toc166076981)

[16 Conclusion 64](#_Toc166076982)

[17 Business Recommendations 64](#_Toc166076983)

[Executive Summary 65](#_Toc166076984)

## **LIST OF FIGURES**

[Figure 1: Overview. 35](#_Toc166028317)

[Figure 2: Missing Values. 40](#_Toc166028318)

[Figure 3: Distribution of Target Variable. 42](#_Toc166028319)

[Figure 4: Distribution of Os 42](#_Toc166028320)

[Figure 5: Distribution of Brands 43](#_Toc166028321)

[Figure 6: Distribution of Networks. 43](#_Toc166028322)

[Figure 7: Distribution of screen size 44](#_Toc166028323)

[Figure 8: Distribution of Rear Camera. 44](#_Toc166028324)

[Figure 9: Distribution of Front Camera. 45](#_Toc166028325)

[Figure 10: Distribution of Internal Memory. 45](#_Toc166028326)

[Figure 11: Distribution of Ram 46](#_Toc166028327)

[Figure 12: Distribution of Battery. 46](#_Toc166028328)

[Figure 13: Distribution of Weight. 47](#_Toc166028329)

[Figure 14: Distribution of Release Year. 47](#_Toc166028330)

[Figure 15: Distribution of Days Used. 48](#_Toc166028331)

[Figure 16: Boxplot for Numeric Attributes. 48](#_Toc166028332)

[Figure 17: Average Price by Ram 49](#_Toc166028333)

[Figure 18: Relation between Used Price and New Price 50](#_Toc166028334)

[Figure 19: Correlation 50](#_Toc166028335)

[Figure 20: Price per Brand. 51](#_Toc166028336)

[Figure 21: Price per OS 51](#_Toc166028337)

[Figure 22: Lasso Models 52](#_Toc166028338)

[Figure 23: Boruta Method 53](#_Toc166028339)

[Figure 24: Tablet- Regression Trees 59](#_Toc166028340)

[Figure 25: Phone- Regression Trees 59](#_Toc166028341)

[Figure 26: Phone- Classification Tree 62](#_Toc166028342)

[Figure 27: Price comparison. 63](#_Toc166028343)

## 

## **LIST OF TABLES**

[Table 1: Attributes 39](#_Toc165795975)

[Table 2: Transforming OS 56](#_Toc165795976)

[Table 3: Transforming Network Type 56](#_Toc165795977)

[Table 4: Transforming Brand 57](#_Toc165795978)

[Table 5: Tablet and Phone Partitioning 58](#_Toc165795979)

[Table 6: Tablet- Regression Models 59](#_Toc165795980)

[Table 7: Phone- Regression Models 60](#_Toc165795981)

[Table 8: Phone Tablet Classification Partitioning 61](#_Toc165795982)

[Table 9: Tablet-train and test records for target 61](#_Toc165795983)

[Table 10: Phone- train and test records for target 61](#_Toc165795984)

[Table 11: Tablet- Classification Models 63](#_Toc165795985)

[Table 12: Phone- Classification Models 63](#_Toc165795986)

# **PROJECT 3: Mortgage Payback Behavior Analytics**

**TABLE OF CONTENTS**

[1. Introduction 67](#_Toc166015223)

[2. Project Overview 67](#_Toc166015224)

[3. Business Understanding 68](#_Toc166015225)

[4. Business Goal 69](#_Toc166015226)

[5. Analytical Approach 69](#_Toc166015227)

[6. Attribute Definition 70](#_Toc166015228)

[7. Data Understanding 72](#_Toc166015229)

[7.1 Number Of Records 72](#_Toc166015230)

[7.2 Checking For Missing Values 72](#_Toc166015231)

[7.3 Checking For 0 Values 73](#_Toc166015232)

[8. Data Preprocessing 75](#_Toc166015233)

[8.1 Optimizing Dataset For Model Efficiency 75](#_Toc166015234)

[8.2 Initial Transformation 75](#_Toc166015235)

[8.2.1 Time Attribute To Observation Count 75](#_Toc166015236)

[8.2.2 Origination Time 75](#_Toc166015237)

[8.2.3 First Time 76](#_Toc166015238)

[8.2.4 Handling Balance Time Column 76](#_Toc166015239)

[8.2.5 Ltv Ratio 76](#_Toc166015240)

[8.2.6 Interest Rate 76](#_Toc166015241)

[8.2.7 Fixed Interest Rate 76](#_Toc166015242)

[8.2.8 House Price Index, Unemployemnt, Gdp 76](#_Toc166015243)

[8.2.9 Real Estate Types 76](#_Toc166015244)

[8.2.10 Investor Type 76](#_Toc166015245)

[8.2.11 Origination Time Columns Handling 77](#_Toc166015246)

[8.2.12 Status Time 77](#_Toc166015247)

[9. Significance Of Analyzing Active Loans 78](#_Toc166015248)

[10. Predictor Analysis 78](#_Toc166015249)

[10.1 Real Estate Types 78](#_Toc166015250)

[10.2 Analyzing Attribute Influence On Default And Payoff 79](#_Toc166015251)

[10.2.1 Investors Status 79](#_Toc166015252)

[10.3 Loan Origination Attributes 79](#_Toc166015253)

[10.3.1 Relation Between Origin Time (Default And Payoff) 79](#_Toc166015254)

[10.3.2 Fico Among Loan Status 80](#_Toc166015255)

[10.3.3 Distribution Of Ltv And Ltv Origin 80](#_Toc166015256)

[10.4 Relation Between Gdp And Unemployment Rate 81](#_Toc166015257)

[10.5 Fixed Interest Rate With Status Time: 82](#_Toc166015258)

[10.6 Economic Indicators For Selected Time Frame 82](#_Toc166015259)

[10.7 Relevancy 83](#_Toc166015260)

[11. Dimension Reduction 83](#_Toc166015261)

[12. Data Transformation 84](#_Toc166015262)

[13. Data Partitioning Methods 84](#_Toc166015263)

[Necessity 84](#_Toc166015264)

[Different Approaches 85](#_Toc166015265)

[Approach In This Project 85](#_Toc166015266)

[Importance 85](#_Toc166015267)

[14. Goal: Classifying Borrowers 85](#_Toc166015268)

[14.1 Model Selections 85](#_Toc166015269)

[14.1.1 Logistic Regression 85](#_Toc166015270)

[14.1.2 Decision Trees 85](#_Toc166015271)

[14.1.3 K-Nearest Neighbors (Knn) 86](#_Toc166015272)

[14.1.4 Naïve Bayes: 86](#_Toc166015273)

[14.2 Model Performance 86](#_Toc166015274)

[14.3 Best Model 86](#_Toc166015275)

[15. Predicting The Status On Active Loans 86](#_Toc166015276)

[16. Conclusion 87](#_Toc166015277)

[17. Business Recommendation 87](#_Toc166015278)

[Executive Summary 88](#_Toc166015279)

## **LIST OF FIGURES**

Figure 1: Overview 1 67

Figure 2: Overview 2 68

Figure 3: Missing values 73

Figure 4: Real estate distribution. 78

Figure 5: Comparison of Investor’s Distribution 79

Figure 6: Analysis of origination time. 80

Figure 7: Distribution of Credit Score 80

Figure 8: Ltv and Ltv origin 81

Figure 9: Relation between GDP and Uer 82

Figure 10: Fixed Interest with status time 82

Figure 11: Correlation Plot 83

## **LIST OF Tables**

Table 1: Attributes 72

Table 2: Missing values for each borrower 73

Table 3: Final Attribute Table 77

Table 5: Duplicate Records 84

Table 6: Data Transformation 84

Table 7: Model Performance 86

PROJECT 1: MAXIMIZING PROFITS FROM SOFTWARE SALES

# **INTRODUCTION**

Initially, a software company, North-Point, sells games and educational software. This company has joined a consortium for customer name pooling. Predictive models will be developed in the project using historical data that focuses on customer response and spending patterns to improve the mailing list. The purpose is to improve the mailing plan for the next software release. In order to maximize the efficiency of models, a sample with an equal set of purchasers and non-purchasers will be used i.e. 1000 each- a total of 2000 observations in the dataset. Two models will be developed where one will categorize the customers into purchasers, and the other model will predict how much they have spent.

Outcome variables:

Purchase – It shows whether a customer purchased a product and responded to a test mailing.

Spending – This attribute comes into the picture if the customer has purchased something. If purchased, then how much they spent to buy that item is spending.

# **PROJECT OVERVIEW**

Figure 1 shows the overview of the project.

A diagram of software production company

Description automatically generated

Figure 1: Overview.

# **BUSINESS UNDERSTANDING**

Companies in the consortium can trade lists of names to a pooled list of customers. Each company adds a list to the pool and can withdraw the same no. Of names every quarter. Predictive modeling is done on the records (in the pool) to improve the selection of customer names. To enhance its mailing approach, north point tested a list including 20,000 names out of a total of 5,000,000 names. 1,065 participants in this test purchased something, generating a 5.3% response rate.

They decided to use 1,000 purchasers and 1,000 non-purchasers to produce a balanced sample for analysis. This simplifies the dataset by producing a noticeable response rate of 50% or 0.5.

By multiplying everyone’s expected probability of purchasing by 0.106, the adjustment brings the dataset's purchase rate back to its initial 5.3%. By doing this, it is ensured that the models are optimized according to the real response rate that was noted during the first mailing test.

# **BUSINESS GOALS**

The objective is to maximize possible gross profit while minimizing mailing expenses. If the company chooses at random from the pool of 180,000 customers, it wants to calculate the gross profit it could expect. Each list booklet costs about $2 to mail, which covers the cost of printing, postage, and other mailing expenses.

The final goal is to pick customers who tend to purchase and spend more to boost the response rate and income from mailing campaigns.

# **ATTRIBUTES DEFINITION**

1. **sequence\_number:** it is the unique identification of each record that is used to refer to records. This attribute might not have any direct impact on prediction.
2. **us:** if a customer is residing in the us the record displays 1 otherwise 0. When developing a model, it will be useful for predictions based on location.
3. "source\_a", "source\_c", "source\_b", "source\_d", "source\_e", "source\_m","source\_o", "source\_h", "source\_r", "source\_s", "source\_t", "source\_u", "source\_p", "source\_x", "source\_w": a customer has been drawn from these sources. This is a binary attribute where 1 indicates a customer name is taken from that particular channel. It will have an impact on customer behavior while making purchases.
4. **freq:** it is a numeric attribute that shows no. Of transactions in the last year. Spending can be predicted and shows how active a customer is in purchasing a product.
5. **last\_update\_days\_ago:** numeric attribute showing how many no. Of days ago the last update was made.
6. **X1st\_update\_days\_ago:** numeric attribute showing how many no. Of days ago the first update was made.
7. **Web.order:** binary attribute that distinguishes customers based on how they make a purchase. If a customer purchased through the web, then 1 otherwise 0.
8. **Gender.male:** gender attribute- if male 1 otherwise 0.
9. **Address\_is\_res:** describe whether the address is the same as the residential address. 1 for residential address and 0 for non-residential address.
10. **purchase:** this is a target variable for classification, which shows whether a customer made a purchase and responded to test mailings.
11. **spending:** this is a target variable for regression that helps maximize gross profit by identifying high-spending customers and offers insights into their purchasing habits.

|  |  |  |  |
| --- | --- | --- | --- |
| Column No. | Variable Name | Description | Type |
| 1 | sequence number | A unique identification number is assigned to each record | Numerical |
| 2 | US | Indicates whether the customer resides in the US (1 for yes, 0 for no) | Binary |
| 3-17 | source\_\* | Binary attributes indicating the customer's source channel which has 15 channels | Binary |
| 18 | Freq | Number of transactions made by the customer in the last year. | Numerical |
| 19 | last\_update\_days\_ago | Number of days since the last update was made. | Numerical |
| 20 | X1st\_update\_days\_ago | Number of days since the first update was made. | Numerical |
| 21 | Web.order | Indicates whether the customer made a purchase through the web (1 for yes, 0 for no) | Binary |
| 22 | Gender.male | Gender of the customer (1 for male, 0 for female) | Binary |
| 23 | Address\_is\_res | Address is same as residential or not (1 for yes, 0 for no) | Binary |
| 24 | Purchase | The customer made purchase in test mailing. | Binary |
| 25 | Spending | Amount in dollars customer spent in test mailing. | Numerical |

Table 1: Attributes

# **DATA UNDERSTANDING**

## **STRUCTURE**

Deciding whether attributes in the north-point project are binary, numerical, or categorical is made easier by understanding the structure.

## **DIMENSION**

Understanding the dataset's dimensions—that is, number of rows and columns—is crucial in evaluating the project's scope.

## **CHECKING FOR MISSING VALUES**

Model predictions may be skewed or incorrect if there are missing values. For, suppose important information is missing then algorithms might not function properly. Complete data provides better modeling in predicting spending and purchaser classification. Checking for missing values can be performed by using is.na () function. This data set does not contain any missing values.

## **CHECKING FOR ZERO VALUES**

For checking 0 values, consider numeric data. This dataset's numerical columns are spending, frequency, last update, and first update. Freq has 398 "0" values, meaning the client did not make any purchases in the previous year. The dataset indicates that there are 1000 non-purchasers. Still, the 0 values are only 999 since one record (sequence no. 711) in the dataset indicates that a customer has not made any purchases and has only paid $1.

## **ATTRIBUTE ANALYSIS**

Analyzing attributes is a crucial step to understanding the features of each variable in the dataset. This involves determining the data types, attribute distribution, identifying outliers, and correlation between the variables.

### **DISTRIBUTION OF OUTCOME VARIABLES**

Purchase: the outcome variable purchase is divided into 2 categories- purchasers and non-purchasers equally as seen below.

Spending: the graph is skewed right where the maximum amount a customer spends is $1500. Most of the people spent below 500 to purchase something.

|  |  |
| --- | --- |
| ***A green and blue rectangular boxes  Description automatically generated*** | ***A purple and white graph  Description automatically generated*** |

Figure 2: Distribution of Outcome Variable

### **DISTRIBUTION OF CATEGORICAL VARIABLES**

We can conclude from the figure 3, that most of the customers are from the us who tend to buy through the web. Gender does not show much of a difference, but male customers are slightly high compared to other genders.

***A screenshot of a graph

Description automatically generated***

Figure 3: Distribution of Categorical Variables.

### **DISTRIBUTION OF NUMERIC ATTRIBUTES**

Frequency and spending are right skewed as seen in figure 4. Customers who made 0,1 transactions are high, so it tops the chart. In the same manner, customers are more likely to spend up to 500 and the maximum amount spent is 1500. Most of the people whose first and last update is around 2500-3000 days ago.

A group of yellow bars

Description automatically generated

Figure 4: Distribution of Numeric Attributes.

### **DISTRIBUTION OF ALL SOURCES WHEN PURCHASED**

The Figure 5, shows the count of each channel or source when purchasing something.

We can see that ‘source a’ has more customers who have purchased and replied to mailing.

***A graph of colored bars

Description automatically generated with medium confidence***

Figure 5: Distribution of Numeric Attributes.

### **ANALYZING THE VARIABLE FREQUENCY WITH OUTCOME VARIABLES**

**Purchase:**

***A graph of a graph showing different colored bars

Description automatically generated with medium confidence***

Figure 6: Distribution of Purchase With Frequency

**Spending:**

A graph of a chart

Description automatically generated with medium confidence

Figure 7: Distribution of Spending With Frequency

Plotted the bar graphs as seen from Figure 6 and 7, to show the relationship between purchase and frequency. From these plots, we could analyze that customers who order once make more purchases.

The scatter plot shows the relationship between the attributes Freq and Spending and it is evident that if the no. of transactions is less the spending is also low. And if the transactions are more customers spend more.

### **AVERAGE SPENDING BY GENDER**

By taking the average of spending, other genders have spent more compared to males.

A green and blue squares

Description automatically generated

Figure 8: Average Spending by Gender

### **BOXPLOTS FOR NUMERIC ATTRIBUTE**

For all numeric attributes boxplot can show the outliers. But, in this case, the outliers do not come into consideration because high spending can show customers who purchased with high amounts rather than errors.

A diagram of a box plot

Description automatically generated

Figure 9: Boxplots for Numeric Attribute.

### **SCATTERPLOT MATRIX TO KNOW THE RELATION BETWEEN EACH NUMERIC ATTRIBUTE**

This entire graph (Figure 10) shows the correlation between all the numeric attributes with their distributions using pairs.panels() function. The last updated and the first updated show a high correlation of 0.81. and the least relation between frequency and spending is noticed.

A screenshot of a graph

Description automatically generated

Figure 10: Scatterplot Matrix

# **PREDICTOR ANALYSIS AND RELEVANCY**

**Sequence numbers** serve just like an index of records which does not give any information that can be used in the models further. This can be considered as an **irrelevant attribute** and removed before building the models for predictions which can actually simplify the dataset. For knowing the important predictors, logistic regression with stepwise can be used which will be performed in the further steps.

There are **90 rows** where all the **source attributes** are marked **0**. Take all those sources with 0 values into **a different set** to analyze. When analyzing this data, records with sources having all 0 values have purchases and spending. It is assumed that these customers are either directly connected with the company or are part of its own database that understands the behavior of customers who interact with the company directly without being taken from any other channel.

# **DATA TRANSFORMATIONS**

The dataset is easier to read when column names are clear and easy to read. The readers can easily understand each attribute's significance without referring to the document. It is simpler to understand the meaning of each attribute if its name is given properly. In this dataset, a few column names can be changed.

|  |  |
| --- | --- |
| Existing column names | New column names |
| Freq | Frequency |
| last\_update\_days\_ago | Days\_since\_Last\_Update |
| X1st\_update\_days\_ago’ | Days\_since\_First\_Update |
| Web.order’ | online\_order |
| Gender.male | gender\_male |

Table 2: New Column Names

# **DIMENSION REDUCTION**

The goal of dimension reduction is to keep the important information in a dataset while reducing the number of variables. This can be advantageous for making the model simpler, lowering the chance of overfitting, and increasing computational effectiveness.

**Correlation analysis:**

This can be performed on numeric data by using summary statistics to identify the highly correlated variables and remove them. The last update and first update as seen in the matrix are near 1 which is highly correlated and overlaps the information, so we can remove anyone but after performing **stepwise** the comparison and reduction can be done if needed.

Mostly the predictors that do not contribute much to the model can be known after building it. So, we can remove the predictors then and evaluate the performances for both models from which we can compare the performances.

After building the model **stepwise** to select important predictors, will come back to this step and make the necessary changes if it is required to be removed.

# **BUSINESS CONSIDERATIONS**

To improve its marketing activities, the company uses tools from financial analysis and predictive modeling. Predictive models are created by analyzing previous customer responses and spending data to enhance the process of choosing names for mailing campaigns. Using these models, the business can improve response rates, more precisely target its customer base, and maximize its marketing budget.

To guarantee precise modeling and analysis, North-Point generates a balanced sample dataset with an equal proportion of buyers and non-buyers. When dealing with uneven response rates, in particular, this balanced sample makes the dataset easier to understand and allows for more precise forecasts.

To guarantee that prediction models accurately represent real-world circumstances, modifications are implemented to the dataset to preserve the initial response rate recorded during test mailings.

The current budget is adequate to meet the project's needs, eliminating the necessity for additional resources currently.

* *Note - Reduction using models:*

Dimension reduction is considered by important attributes in the dataset. When building the model, logistic regression can be used for this purpose. If the model is performing better with the important predictors – only those can be taken into consideration and suppose there is no improvement, then can consider all the predictors. In that case, dimension reduction will not take place. Not every dataset needs to perform dimension reduction. Based on the insights, whatever is needed can be done.

# **DATA PARTITIONING**

## **WHY IS DATA PARTITIONING NEEDED?**

Data partitioning is important because it evaluates model performance on data that is not used in the training set. This guarantees that the model applies effectively to the new set of data. Model performance can be improved without overfitting. Overfitting is avoided by dividing the data or splitting it. Data partitioning is an essential step to ensure that models are reliable and able to produce correct predictions.

## **METHODS**

Holdout Method: This is splitting the data into 2 or 3 parts training and testing or training, validation, and testing. This works by building the model on training, if validation is considered then predict the model on validation and evaluate different model performances which is called as fine-tuning, and then test on completely new data with the best model from validation.

Cross-Validation: The dataset is divided into folds. The model will be trained on selected folds like k-1 folds and tested on other folds.

To achieve effective model performance and interpretability in the classification process, logistic regression with backward selection will be used to check which model is performing better.

As a part of the project requirement data partitioning is done this way because it serves the purpose:

Training with 800 records

Validation with 700 records

Testing with 500 records

## **IMPORTANCE OF TRAINING, VALIDATION AND TESTING**

Its main purpose is before the final assessment on the test data, it optimizes the model on validation data. The model is guaranteed to be trained on a training set, validated to adjust parameters and prevent overfitting (validation set), and then select the best model from validation, then predict the selected-on test data.

After partitioning the data, models should be selected for classifying a customer into purchaser or non-purchaser & to predict the spending when a purchase is made. The best model should be selected by comparing the performance of different models based on evaluation metrics. Training the model, evaluating, comparing, and selecting the best will be seen in further steps.

# **MODELS SELECTION**

## **GOAL 1- CLASSIFICATION**

The main goal as discussed above is to first classify a customer into a purchaser or non-purchaser. This enables companies to target those who are most likely to react favorably to campaigns, promotions, and mailing. This can be achieved by performing logistic regression and with the backward method as per the requirements. Although other possible models can also be built for just comparing purposes.

As we are aware the company is involved in mailing campaigns, classification models pick customers who respond to the campaign, which helps with mailing list efficiency and raises response rates while lowering mailing costs. This step is essential for keeping current customers as well as for attracting new ones. Spending can be removed in this classification.

**To classify the customer into a purchaser or non-purchaser these models can be used:**

### **LOGISTIC REGRESSION**

This glm() model is well suited for binary classification problems that give probabilities of a customer belonging to a particular class.

Initially consider all the predictors and build the model on a train set, then make predictions on the validation data. It gives predicted probabilities of purchase for each observation in the validation dataset. The probabilities are continuous values between 0 and 1. But to classify a customer it should be binary like 0- for non-purchasers and 1 for purchasers. So, to convert these probabilities into binary values threshold is used. The threshold is 0.5. If the probabilities are greater than the threshold- classified as 1(purchaser) less than the threshold- classified as 0(non-purchaser). Binary labels should be factored, and levels should be set to “1” and “0”. By doing so the prediction will be appropriate.

### **BACKWARD METHOD**

To improve the model performance other methods like selecting important predictors can be taken into consideration and checking if there is any improvement. StepAIC will be available in the MASS package.

This method gives the most relevant and significant variables for predicting whether a customer is a purchaser or non-purchaser. By identifying and eliminating strongly correlated variables, stepwise techniques can lessen multicollinearity problems that could compromise the model's stability and interpretability.

Mostly backward stepwise is preferable, so for this model backward is used.

After performing stepwise on the initial model the predictors, it is considering important are*:*

*source\_a, source\_e, source\_h , source\_r , source\_s, source\_t, source\_u, source\_p, source\_x, source\_w, Frequency, Days\_since\_last\_update, online\_order, Address\_is\_res*

In the earlier steps, dimension reduction shows a high correlation between the last update and the first update where it is mentioned that anyone can be removed. From the important predictors, we can see that the first update is removed as it is considered irrelevant. Model evaluation should be for every model to conclude which model is performing well and justify it.

For this model and the model with all predictors, there is not much difference but a slight decrease in sensitivity and accuracy in model 1 is noticed.

**Comparing Logistic Regression models based on accuracy and sensitivity***:*

After performing the model with stepwise and all predictors, can check by removing insignificant variables in each step. By doing so if there is any improvement in the model performance on the validation data we can select that model to predict on holdout data.

Model 2: source\_s amd source\_x are insignificant so these variables are improved in this model.

Model 3: source\_t is insignificant so removing that in this model.

Model 4: source\_e and last\_updaate\_days\_ago are less significant marked as ‘.’, removing them.

Model 5: source\_p is insignificant so removing that in this model.

Model 6: source\_r has just one star so trying by removing it and checking the performance.

|  |  |  |
| --- | --- | --- |
| Logistic Regression Model | Accuracy | Sensitivity |
| Initial Model | 79.57% | 77.59% |
| Model 1(After stepwise) | 79.29% | 76.44% |
| Model 2 (Significant variables set 1) | 79.29% | 76.15% |
| Model 3 (Significant variables set 2) | 78.86% | 74.71% |
| Model 4 (Significant variables set 3) | 79.71% | 75% |
| Model 5 (Significant variables set 4) | 79.57% | 74.71% |
| Model 6 (Significant variables set 5) | 78.57% | 72.99% |

Table 3: Logistic Regression Models

Model 4 seems to have the highest accuracy compared to all other models, but the sensitivity is just 75%. Our target is to pick the best customers for the firm, which means classifying as a purchaser is important so the sensitivity shows how likely a customer can be classified as a purchaser correctly. By considering that fact we must go with more sensitivity so an initial model with all predictors can be considered from all these models.

### **CLASSIFICATION TREE**

The classification tree is easy to understand and gives a simple way to demonstrate decision rules. Each node represents a decision. When it comes to predicting the target variable, it chooses the important predictors by itself. By this, we can easily identify the important attributes that impact the classification. The leaf nodes will have 0 or 1 showing non-purchasers or purchasers, predictor and value of the split will be shown right below the node. The count of the classes will be shown inside the node.

The training dataset builds the decision tree based on the partition. Another set of 700 records which is not used in training is used to check the performance- notices for overfitting where they do not perform well on new data.

To compare the tree - important predictors are also taken into consideration to check the performance. The process is the same as the above. Here when considering the important predictors model tends to perform well compared to another tree model. Tree is also **pruned** for better results.

**Tree after pruning with all predictors:**

A diagram of a computer

Description automatically generated

Figure 11: Classification Tree with all Predictors.

**Tree after pruning using important predictors:**

A diagram of a computer network

Description automatically generated

Figure 12: Classification Tree with Important Predictors.

### **K-NEAREST NEIGHBOR**

Problems involving binary and multi-class classification can be handled by k-NN. It is easy to apply k-NN to the North Point dataset, where we classify customers into purchasers and non-purchasers. Based on the majority class of the k nearest neighbors to a particular data point, KNN generates predictions.

**KNN with all predictors:**

The model is trained. Next predictions are made using the model on the validation dataset. Following the prediction- accuracy and sensitivity are considered for the model performance. Validation data reveals a sensitivity of 76.15.

**KNN with important predictors from the backward method:**

A set of predictors is taken from the backward method performed in the logistic regression to compare the performances. Here, the goal is to determine which model performs better- only significant predictors in a model or by using all the predictors.

**COMPARE THE CLASSIFICATION MODELS WITH ACCURACY AND SENSITIVITY**

|  |  |  |
| --- | --- | --- |
| Models | Accuracy | Sensitivity |
| Logistic Regression model with all predictors | 79.57% | 77.59% |
| Logistic Regression model with backward | 79.29% | 76.44% |
| Classification Tree | 76.86% | 67.53% |
| Classification Tree (with important predictors) | 77.5% | 75.86% |
| KNN (k=1) | 80.71% | 76.15% |
| KNN (with important predictors) | 78.86% | 77.30% |

Table 4: Classification Model Selection

The classification tree has the least sensitivity so that model can be eliminated, considering Logistic regression with all predictors to perform on hold-out data as the sensitivity is more for that model and meets the project requirements.

**Logistic regression with all predictors can be considered to perform predictions on the final set which is the test dataset.**

## **GOAL 2- REGRESSION**

The goal is to project the amount of money that customers will likely spend on their purchases. With the aid of this projection, North Point will be able to precisely project income and customize marketing tactics to optimize profitability***.***

**To predict spending value for purchasers these models can be used*:***

### **LINEAR WITH STEPWISE REGRESSION**

Linear regression is used to predict numeric outcomes. This data predicts the spending should contain only purchasers which means Purchasers = 1 should be considered. Later, the purchase should be removed while building the model.

Based on the most influential predictors, they can make marketing strategies. This is a fundamental model for predicting spending. A useful way to verify the validity and reliability of a model is to compare its performance with sophisticated models like regression trees.

The first step is to build a multiple linear regression model including all predictors. Build a model on the train data and evaluate using the validation set. Now, perform stepAIC with backward selection to get the important predictors. Once we have the important predictors, repeat the same steps and compare the results. This model will be using mean absolute error for evaluation. The least mae shows better performance in the model.

Comparing both the linear regression models, MAE was low for the backward method (109).

So, let's dive into exploring regression trees for comparison.

### **REGRESSION TREES**

The regression tree helps in predicting spending values for each purchaser. We must first identify predictors and later split the data. The model is built on training data. Tree splits based on the highest standard deviation reduction (SDR).

The tree automatically takes the important predictors into consideration so build the tree on the train and evaluate the mae score on validation data.

A diagram of a number

Description automatically generated

Figure 13: Regression Tree

### **COMPARING MODELS FOR PREDICTING SPENDING**

|  |  |
| --- | --- |
| Models | Evaluation (MAE) |
| Multiple linear regression with all predictors | 111.902 |
| Multiple linear regression after stepwise | 109.8421 |
| Regression tree | 100.581 |

Table 5: Comparing Models for Predicting Spending.

**Linear regression after stepwise can be considered as it deals with important predictors only that have low MAE which helps to deal with model complexity, also it is a part of the specifications mentioned.**

## **MODELS SELECTED TO PREDICT ON HOLDOUT DATA**

* Classification - Logistic regression with all predictors
* Regression - Linear regression using backward.

# **HOLDOUT DATA ANALYSIS AND ADJUSTMENTS:**

It's crucial to examine the original holdout data and make the required modifications for more precise predictions after developing the selected models and assessing their effectiveness. By doing this step, we can be sure that the models are realistic and optimized for real-world situations.

After predicting the selected models on test data these are the results:

Evaluation of test data for both selected models shows improved performance.

## **COMPARING LOGISTIC REGRESSION ON VALIDATION AND TEST DATA**

|  |  |  |
| --- | --- | --- |
|  | ACCURACY | SENSITIVITY |
| VALIDATION DATA | 79.57% | 77.59% |
| TEST DATA | 82.6% | 79% |

Table 6: Comparing Logistic Regression on Validation and Test Data.

## **COMPARING LINEAR REGRESSION ON VALIDATION AND TEST DATA**

|  |  |
| --- | --- |
|  | MAE (Mean Absolute Error) |
| VALIDATION DATA | 109 |
| TEST DATA | 113 |

Table 7: Comparing Linear Regression with Backward on Validation and Test Data.

# **PROFIT ANALYSIS**

North Point's business goals include maximizing gross profit, reducing mailing costs, and successfully targeting customers to increase response rates and revenue from mailing campaigns can be done by adding the following columns:

## **GROSS PROFIT**

North Point can more accurately estimate and plan its financial resources by estimating the possible gross profit. It assists in overall financial planning for marketing efforts and will know where it requires improvement. The overall goal is to maximize the profits.

Gross Profit can be calculated by considering:

* Total no. of customer in mailing pool = 180,000
* Response Rate = 0.053 (5.3%)
* Average of spending when purchase is 1 = 205
* Each mailing cost = $2

By multiplying the response rate (0.053) by the average spending of the customers who made a purchase and deducting the mailing cost per customer ($2). This gives an approximate idea of how much money North-Point Software Production Company might make from the campaign. Estimated gross profit is $1598075.

## **PREDICTED PROBABILITY OF PURCHASE**

Based on the predictive logistic regression model on test data, this column assists North Point in determining the probability that each customer will make a purchase. It enables the business to provide preference to clients who are more likely to react favorably to marketing initiatives, which improves resource allocation and boosts conversion rates.

## **ADJUSTED PROBABILITY OF PURCHASE**

The Adjusted Probability of Purchase column adjusts for purchaser oversampling, guaranteeing that the predictive models correctly represent the actual response rate seen in the first mailing test. North Point can produce more accurate forecasts and adjust marketing initiatives by aligning the dataset with the real purchase distribution. This column can be added by multiplying the predicted probability of purchase by the original purchase rate which is 0.1065.

## **PREDICTED SPENDING VALUE**

North Point may more precisely predict revenue and customize marketing tactics to increase profitability by projecting the amount of money that customers are likely to spend on purchases. Personalized offers and targeted marketing efforts boost client engagement and loyalty by considering the spending patterns of various customer categories.

## **EXPECTED SPENDING**

The expected spending column offers a more accurate estimate of customer’s expenses. With this modification, oversampling is taken into consideration and accurate projections are used to provide revenue forecasts. This allows North Point to make well-informed judgments.

## **CUMULATIVE GAIN CHART**

The cumulative gain chart seen in Figure 14, provides information on the efficiency of targeting methods by visualizing cumulative predicted spending as a function of records targeted. It assists North Point in determining high-value customer segments, assessing the return on investment of various marketing strategies, and optimizing campaign effectiveness to increase revenue production. The chart shows selecting around 400 customers from the test records will maximize the spending which increases the profit.

A graph and a graph

Description automatically generated

Figure 14: Cumulative gain chart.

# **CONCLUSION**

North-Point successfully implemented predictive modeling to optimize its mailing campaigns, resulting in improved response rates and increased profitability. The adjusted dataset ensured realistic modeling, while financial analysis tools aided in profit estimation and resource allocation. Overall, North-Point's strategic approach led to enhanced marketing effectiveness and maximized gross profit, validating the project's objectives and methodologies.

# **BUSINESS RECOMMENDATIONS**

North Point must constantly assess how well its marketing initiatives and prediction models are working and modify its approach as necessary. North Point can preserve a competitive edge and be successful in its marketing efforts over the long run by remaining adaptable and sensitive to shifts in customer behavior and market dynamics. Using the cumulative chart, North Point should constantly analyze the success of its targeting methods to pinpoint high value.

# **EXECUTIVE SUMMARY**

**NAME:** Sreeja Reddy Singidi

**DATE:** 05-08-2024

**OPPORTUNITIES:**

The project provided valuable insights into customer behavior and spending patterns, enabling North-Point to better understand its audience and tailor its offerings to meet their needs. This deeper understanding facilitated more personalized marketing strategies, driving increased customer satisfaction and loyalty.

**SOLUTIONS:**

The company tried to improve its mailing strategy for games and educational software by utilizing financial research and predictive modeling. Additionally, throughout the project, significant attention was paid to business factors like optimizing gross profit, reducing mailing costs, and successfully targeting customers. More informed decisions and resource allocation were made possible by the inclusion of modified probabilities, spending forecasts, and predicted purchase probabilities.

PROJECT 2: EMPOWERING USED MOBILE - DEVICE BUYERS

# **INTRODUCTION**

Used smartphone demand has increased significantly in today's dynamic world of technology and business, driven by several results like affordability, sustainability, and technological improvements, but it's still quite difficult to determine these device’s true worth among the many different features, and brands. "Predict & Win, is a gaming platform that aims to tackle this difficulty by providing consumers with an organized approach to understanding the complicated world of pricing for used smartphones. It also provides engaging experience in estimating the value of used devices. Within this innovative environment, users explore the complex dynamics of the used mobile phone market in a creative environment, taking them on an interesting trip.

The platform will include an easy-to-use interface with simple controls for submitting predictions and monitoring results.

The gamification elements also encourage active involvement and skill improvement by rewarding users with points and badges which they can redeem when purchasing a used device.

# **PROJECT OVERVIEW**

The overview of the project can be seen in Figure 1.

A diagram of a diagram

Description automatically generated with medium confidence

Figure 1: Overview.

# **BUSINESS UNDERSTANDING**

Predict and Win operates as an interactive gaming platform where users engage in estimating the resale price of used mobile phones. Additionally, the platform serves as a marketplace where users can purchase used devices, thereby offering a comprehensive solution for both gaming and e-commerce needs.

As an interactive gaming platform, it allows players to estimate the cost of reselling used mobile phones. To generate meaningful estimates, participants examine a variety of elements using an easy-to-use interface, including device characteristics, market trends, and historical data. These rough estimates provide the foundation for earning rewards that can be redeemed while purchasing a device, which increases user motivation and engagement.

Predictive analytics is applied in "Predict and Win" to create models that reliably forecast used mobile phone resale values. To guarantee that customers are presented with realistic estimation scenarios, these models are continuously improved through the application of sophisticated algorithms and machine learning approaches.

# **BUSINESS GOAL**

The primary business goal of "Predict and Win" revolves around revenue generation. Premium challenges can be purchased using participation fees as one source of revenue. To participate in high-stakes challenges with exclusive rewards, users may be required to pay an entry fee, which goes toward funding the platform. For customers who are prepared to pay a participation fee, this will provide premium challenges with greater stakes and exclusive rewards. These challenges make money through entrance fees and sponsorships, offering devices, or sponsored gifts from collaborating firms.

Additional revenue opportunities can be obtained through partnerships with industry partners, merchants, and device manufacturers. Through revenue-sharing agreements or sponsorship payments, these partners may support contests, feature new items, and raise brand recognition on the site.

# **ANALYTICAL APPROACH**

The analytical approach to build models that can be used to predict the used prices for the game includes the following:

Data analysis: The data should be first cleaned- handling the missing values and zero values. This part includes attribute analysis like the distribution of target variables, discovering patterns, finding outliers, and knowing the relationship between each variable.

Model building: We will be using classification and regression algorithms for used price categories and used price predictions respectively.

Model evaluation: In further steps, the best model will be picked- its accuracy and performance will be assessed based on that.

Deployment: After the testing is done, they can be deployed for used price recommendations and categories for used devices.

After constructing predictive models for estimating used smartphone prices and categories, these predictions become the backbone of the gaming experience. Customer’s predictions are based on these model-estimated predictions, and they receive prizes if their guess is within the expected range; if not, no additional points are awarded.

By providing rewards for precise prediction, this approach raises user involvement and promotes data-driven decision-making. It guarantees that the prediction models have a useful function within the game platform, providing users with an enjoyable and enlightening experience.

# **ATTRIBUTE DEFINITION**

1. device\_brand: This attribute shows the brand name of the smartphone like Apple, Infinix, Huawei, etc. This will help to know the manufacturer of the smartphone and also for any analysis that is brand specific. The type of the variable is character. ‘Others’ means brands that are not well-known or comparatively smaller brands.
2. os: This attribute tells which operating system is used in the phone like IOS, Android, Windows, or others which are not used much.
3. screen\_size: The attribute shows the screen size of the smartphones in inches which is a numeric type. It helps identify the physical dimensions of the display, which can improve the customer’s usability.
4. X4g: This is a binary attribute that shows whether the given smartphone supports a 4G network or not. If it supports 4g then yes, otherwise no.
5. X5g: This is similar to the above attribute which shows whether the given smartphone supports a 5G network or not. If it supports then yes, otherwise no.
6. rear\_camera\_mp: This is an important feature of the smartphone for youth as they focus on taking good pictures and videos. This shows how many megapixels the main camera which is on the back has. Megapixel just means how much detail the camera can capture. More the megapixel more the clarity the photos.
7. front\_camera\_mp: It is same as the back camera. This concentrates on the megapixels of the front-facing camera in smartphones which shows how well the front camera can capture selfies or self-portraits.
8. internal\_memory: It describes the amount of storage that can be used to store files, media, and applications that is numeric. It affects the cost, usability, and general performance of the phone. Customers frequently take internal memory size into account to know if a used smartphone is valuable and appropriate for their needs.
9. ram: Used smartphone must have Random Access Memory because it is essential for multitasking and temporary data storage. In this case - larger RAM capacities are preferable since they can perform multiple activities at once without experiencing any lag.
10. battery: The battery is measured in mAh (milliampere-hours) which tells how much power the phone can hold. For customers who use the phone all day, battery becomes essential. An increased battery life is linked with higher mAh values.
11. weight: Shows the weight of the phone in grams.
12. release\_year: Understanding the smartphone's release year affects its value in the market by allowing one to look at its technological innovations. To help with pricing and estimates of demand for used phones, the release year analyzes the trends and the understanding of customer preferences over time.
13. days\_used: This attribute shows the number of days the smartphone has been in use since purchase which may have an impact on its state. Provides insights into customer behavior related to the lifespan and replacement cycles of devices.
14. normalized\_used\_price: For determining the cost of used smartphones based on their characteristics and specifications is made by standardizing the values. Moreover, the models can successfully identify trends and patterns in the used smartphone market by using normalized prices, which helps buyers and sellers make smarter decisions.
15. normalized\_new\_price: These are the standardized values of the smartphone's original price based on the features. It adjusts the prices to the same range, to give fair and clear comparisons across different models and brands.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Type |
| device\_brand | Brand of the smartphone | Categorical |
| os | Operating system of the smartphone | Categorical |
| screen\_size | Size of the smartphone screen in cms | Numeric |
| X4g | Phone having 4g or no | Categorical (yes or no) |
| X5g | Phone having 5g or no | Categorical (yes or no) |
| rear\_camera\_mp | rear camera in megapixels | Numeric |
| front\_camera\_mp | front-facing camera in megapixels | Numeric |
| internal\_memory | Internal memory capacity of the smartphone in GB | Numeric |
| ram | Random Access Memory (RAM) of the smartphone in GB | Numeric |
| battery | The capacity of the smartphone's battery in milliampere-hours (mAh) | Numeric |
| weight | Weight of the smartphone in grams | Numeric |
| release\_year | Year of release of the smartphone | Numeric (Year) |
| days\_used | Number of days the smartphone has been used | Numeric |
| normalized\_used\_price | Normalized price of the used smartphone | Numeric |
| normalized\_new\_price | Normalized price of the original smartphone | Numeric |

Table 1: Attributes

# **DATA UNDERSTANDING**

The dataset consists of 3454 records with 15 variables that are the features of the used phone with the target variable.

## **SUMMARY STATISTICS**

Summary statistics will be relevant for numeric data because they show the mean, median, and quartiles of the variables with min-max values.

***A screenshot of a computer screen

Description automatically generated***

Code Snippet 1: Summary Statistics

## **CHECKING FOR MISSING VALUES**

There are 202 missing values in total and a huge number of missing values in the variable rear camera\_mp which is the back camera as seen in Figure 2.

A graph with blue squares and pink squares

Description automatically generated

Figure 2: Missing Values.

Investigating the highest missing values variable i.e, rear camera:

Missing values in rear camera relation with year:

When going through the dataset we can notice that 122 records are missing from the *year 2022* out of 277 records in that particular year. This might be because they did not record these details consistently or brands might not have focused much on giving details about the latest phone’s rear cameras.

Relation with Operating System:

All 179 records are associated with Android phones.

**Replacing missing values by group medians:**

The dataset is initially grouped by brand, creating separate groups for each unique brand. Within each group, the median value for specific columns is calculated to handle missing values.

However, even after this process, 10 records remain missing. Upon investigation, it's discovered that all these missing records belong to the brand "Infinix." Infinix models from the years 2017 to 2020 are known to have rear cameras. This discrepancy suggests that the missing values for the rear camera are likely due to data entry errors or incomplete data for Infinix models. Given this, these 10 records are considered unreliable and can be safely removed from the dataset.

## **CHECKING FOR 0 VALUES**

Only the front camera has 39 zero values as seen in Code snippet 2.

Analyzing the reason for these 0 values - all 0 values are from the Nokia brand that supports ‘Other’ operating systems apart from Android and Windows with no 5g network. The reason could be that fewer common models did not prioritize the front camera whereas the manufacturers might have designed the mobile for other specific features rather than the front camera.

Nokia had its own Symbian OS but later shifted to Android and Windows due to external factors. So, this can be one of the reasons for these 0 values or the records might be missing. For now, those zero values need not be changed.

A close-up of a computer code

Description automatically generated

Code Snippet 2: Zero Values.

## **ATTRIBUTE ANALYSIS**

### **DISTRIBUTION OF TARGET VARIABLE**

Blue to red color in Figure 3, shows the relative frequency of used device prices inside each bar, which reflects a certain range of normalized values. Bars with red color show areas with higher concentrations of used phones. The highest prices go between 4-5 from the plot below. The shape shows the distribution of prices.

A graph of a normalized used price

Description automatically generated

Figure 3: Distribution of Target Variable.

### **CATEGORICAL ATTRIBUTES**

#### **OPERATING SYSTEM**

According to Figure 4, Android has the highest count of used smartphones than others. This shows that this dataset sees higher sales in Android due to the availability of different used device prices and because of customer preferences.

A graph of operating system

Description automatically generated

Figure 4: Distribution of Os

#### **DISTRIBUTION OF BRANDS**

We can analyze the features and pricing of top brands. By doing this it is possible to compare the best brands to the others. We can also analyze predictive models that focus on top brands to check if they can improve the accuracy of models. The top brands from Figure 5, are Others (brands apart from this dataset), Samsung, Huawei, LG, and Lenovo.

A graph of different colored bars

Description automatically generated

Figure 5: Distribution of Brands

#### **NETWORKS TYPE**

From Figure 6 it indicates the majority of smartphones support 4g. We can say that 5G is more recent and has not yet become as popular in the used smartphone market. By identifying instances where devices lack 4G and 5G capabilities, we gain an understanding of the distribution of network technologies in the market. There are higher chances of buyers considering “other” networks like 3g, and 2g than 5g.

A graph of a network distribution

Description automatically generated

Figure 6: Distribution of Networks.

### **NUMERICAL ATTRIBUTES**

#### **DISTRIBUTION OF SCREEN SIZE**

The distribution of screen sizes indicates that 12-15 inches is the range in which most mobile devices fall, indicating a widely accepted standard size in the market. With a screen size of thirty inches as seen in Figure 7, there is an interesting outlier that suggests a gadget that probably might be a tablet.

*A graph of green bars

Description automatically generated*

Figure 7: Distribution of screen size

#### **DISTRIBUTION OF REAR CAMERA**

The examination of back camera megapixels indicates a widespread tendency toward 13-megapixel cameras. But significant percentages also include 6 and 8-megapixel cameras, showing variation in camera specifications. In this dataset, no phone has megapixels between 25 to 40 possibly indicating a lack of devices with higher-resolution cameras with this dataset.

A graph of a camera size

Description automatically generated

Figure 8: Distribution of Rear Camera.

#### **DISTRIBUTION OF FRONT CAMERA**

Notably, 6–7-megapixel front cameras are the most popular option, indicating possible availability in this market. Furthermore, the existence of phones with front cameras over 30 megapixels suggests that there is a wide range of options available to suit customers' demands for different levels of performance and quality. These kinds of insights are crucial for buyers and sellers to successfully navigate the used phone market.

A graph of a camera size

Description automatically generated

Figure 9: Distribution of Front Camera.

#### **DISTRIBUTION OF INTERNAL MEMORY**

The majority of used smartphones in the dataset possess internal memory capacities below 100GB, indicating common storage configurations. However, the presence of a record with 1000GB of internal memory suggests an outlier observation.

A graph with green and black bars

Description automatically generated

Figure 10: Distribution of Internal Memory.

#### **DISTRIBUTION OF RAM**

According to an analysis of the RAM distribution, only a small percentage of smartphones have 8GB of RAM, while the vast majority of devices have 4 GB. This finding shows that consumers strongly prefer smartphones with 4GB of RAM.

*A graph with green rectangular bars

Description automatically generated*

Figure 11: Distribution of Ram

#### **DISTRIBUTION OF BATTERY**

Only a small percentage of smartphones have battery capacities larger than 6000 mAh. Most used smartphones have battery capacities between 2000 and 4500 mAh. This finding highlights the widespread practice of smartphones having batteries that fall within a specific capacity range, which is important information for buyers and sellers on the trade-in platform to know when evaluating aspects like battery life and performance.

*A graph of battery

Description automatically generated*

Figure 12: Distribution of Battery.

#### **DISTRIBUTION OF WEIGHT**

It can be seen from Figure 13, that a large percentage of the smartphones in the dataset weigh between 140 and 190 grams and a phone has 850 grams as the highest value.

*A graph of a weight

Description automatically generated*

Figure 13: Distribution of Weight.

#### **DISTRIBUTION OF RELEASE YEAR**

The years 2013 and 2014 show the highest frequency of smartphone purchases, indicating an increase in demand from consumers during this time. This trend may be explained by the notable developments in smartphone technology over the past several years.

The least number of purchases, however, was made in 2020, suggesting a possible drop in customer interest or market saturation. This decline in purchases could be because of the **COVID** pandemic.

*A graph of a number of green bars

Description automatically generated with medium confidence*

Figure 14: Distribution of Release Year.

#### **DISTRIBUTION OF DAYS USED**

Most of the phones in the dataset were in use for 600–900 days after purchase, which suggests that consumers often keep their phones for this amount of time. Understanding the distribution of days used can help in assessing the depreciation rate of smartphones over time.

*A graph of green and black bars

Description automatically generated*

Figure 15: Distribution of Days Used.

#### **BOXPLOT FOR NUMERIC ATTRIBUTES**

In the interquartile range (IQR), the median is indicated by the horizontal line inside the box, and the whiskers extend to the minimum and maximum values within 1.5 times the IQR from the first and third quartiles, respectively, for each box plot that shows the distribution of values for a particular attribute. This visualization provides insights into the distribution and variability of numeric attributes in the dataset.

*A screenshot of a computer

Description automatically generated*

Figure 16: Boxplot for Numeric Attributes.

# **PREDICTOR ANALYSIS AND RELEVANCY**

# **ANALYSIS**

### **AVERAGE PRICE BY RAM**

The line graph (Figure 17) shows the relationship between RAM capacity and the average price of used devices. It shows a continuous increase up to a RAM capacity of 12 with slight decreases in the price when RAM was high. This graph is useful for understanding how RAM capacity influences the price of used devices. Taking the average ensures that the graph represents a generalized trend across different RAM capacities rather than individual data points, providing a clearer picture of the overall relationship.

*A graph of different colored bars

Description automatically generated*

Figure 17: Average Price by Ram

### **RELATION BETWEEN THE USED PRICE AND A NEW PRICE**

The scatter plot indicates that they have a positive correlation, which means that as the normalized new price increases, the normalized used price tends to increase. This is to be expected because phones with higher new prices will probably also cost more when used. Knowing how new and used prices relate to one another can help customers decide which one to buy. They have a better chance to decide whether to buy new or used phones depending on their financial restrictions since they can predict how a smartphone's resale value may fluctuate over time.

*A graph showing a normalized price

Description automatically generated*

Figure 18: Relation between Used Price and New Price

### **CORRELATION BETWEEN THE NUMERIC DATA**

Strong correlations were found between predictors, such as battery weight and weight-screen size, suggesting possible relationships between these features. On the other hand, internal memory-screen size and rear camera-5G showed lower correlations, indicating a weaker relationship between these factors. By understanding these correlation patterns, model-building procedures can be improved.

*A diagram of a device

Description automatically generated with medium confidence*

Figure 19: Correlation

### **PRICE PER BRAND**

The dataset contains a roughly uniform representation of smartphones across various price ranges which means there isn't a significant skew towards higher or lower-priced devices, and smartphones are distributed relatively evenly across different price categories.

*A graph of a company brand

Description automatically generated*

Figure 20: Price per Brand.

### **PRICE PER OS**

The graph (Figure 21) suggests that iOS devices have a higher value in the market, likely due to factors such as brand reputation. On the other hand, Android devices, while still popular, tend to be priced lower than iOS devices, reflecting their broader availability across a range of price points and manufacturers. Other operating systems, which are less widely used by the public, have minimal representation in the market and consequently do not attract significant demand or make high prices.

*A graph of a bar chart

Description automatically generated with medium confidence*

Figure 21: Price per OS

# **RELEVANCY**

To build precise predictive models, it is essential to identify and rank the most significant predictors. This can be performed by *feature selection*.

### **LASSO MODEL**

LASSO will consider irrelevant predictor’s coefficients to exactly zero to show variable selection. By eliminating predictors from the model that have little to no impact on the used price, LASSO effectively modifies the data. The predictors whose coefficients are non-zero after regularization are regarded as significant and are kept in the model. This model automatically considers all categorical attributes into dummies.

It is concept- oriented which gives better results for feature selection by shrinking the errors. So, this is considered an important method.

The significant predictors using this method are:

**

Code Snippet 3: Lasso Important Predictors.

Using cv.glmnet: Using glmnet:

|  |  |
| --- | --- |
| **A graph with numbers and a red dotted line  Description automatically generated**  Figure 22: Lasso Models | A graph of a number  Description automatically generated with medium confidence |

### **BORUTA METHOD**

A wrapper-based feature selection technique called Boruta analyzes each predictor's significance. To determine their significance iteratively balance the weights of shadow attributes (random noise) to real features. In the end, Boruta classifies features as "Confirmed," "Tentative," or "Rejected" according to how important they are.

Every predictor appears to be significant from Figure 23, according to the Boruta feature selection plot. We'll use the random forest to verify the significance of this finding.

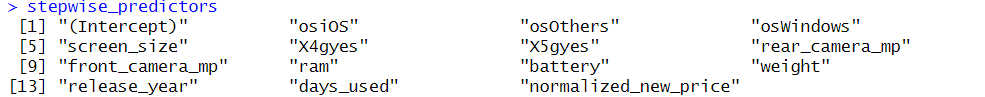
**A graph with green and blue squares

Description automatically generated**

Figure 23: Boruta Method

### **STEPWISE REGRESSION**

It is a statistical technique where predictors are gradually added or removed accordingly.

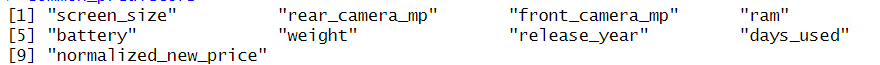


Code Snippet 4: Stepwise Regression Predictors.

The common predictors found in all models after using these feature selection techniques are believed to be the most important variables for predicting the normalized used price of smartphones.

Intersecting the predictors with these models can help identify the important common features which can be seen in Code snippet 5.

These common features are likely to have a significant impact:



Code Snippet 5: Common significant predictors.

However, the final decision on which predictors to include in the models will be based on thorough evaluation during model building. Both Lasso and stepwise have similar important predictors. We'll compare models built with all predictors against those built with the identified important predictors using stepwise to make correct decisions about feature selection and model performance.

# **DATA TRANSFORMATION**

## **CHANGING COLUMN NAME**

The data transformation phase involves modifications to enhance readability and usability. The column name “device\_brand” can be changed to “brand”.

# **DIMENSION REDUCTION**

## **COMBINING NETWORK TYPE**

The 4G and 5G network separate columns are combined into a single column called "network\_type," which simplifies the dataset. It is a well-established concept that devices with 5G connections also have 4G connectivity by default. In addition to the differentiations between 4G and 5G networks, there is also a further category called "Other." This group includes situations in which a network does not support 4G or 5G and hence falls under the "Other" category.

# **DATA PARTITIONING METHODS**

**NECESSITY**

Data partitioning is essential to predict models accurately. By dividing the dataset, we can train models on a portion of the data, assess their efficiency, and ensure the models operate well when applied to new, untested data. By preventing overfitting, optimizing model performance, and offering trustworthy smartphone pricing predictions, this method improves the effectiveness and reliability of the gaming platform.

**DIFFERENT APPROACHES**

Data partitioning can be done in different methods:

**Train-test division**: The process of train-test partitioning involves splitting the dataset into two subsets: a training set for model training, and a testing set for model evaluation. This is expanded by train-valid-test partitioning, which includes a validation set for optimizing model parameters and hyperparameters for improved generalization and performance on new data which is divided.

**Cross Validation**: It is a resampling strategy that divides the dataset into several subsets of training and validation sets iteratively to assess the performance of the models. Training and testing a model on several subsets of the data lowers the likelihood of overfitting and offers a more reliable evaluation of the model's performance, which shows how well the model will generalize to new data.

**APPROACH IN THIS PROJECT**

 The straightforward train-test split approach will be used in this project, allocating **70%** of the dataset for training and **30%** for testing. There is a significant amount of data to train the models, with 70% of the data. This gives the models enough information to identify patterns and connections in the data. Considering 30% of the data for testing ensures that we have a sufficient dataset to evaluate the model's performance accurately. This helps assess how well the trained models generalize to new data that is test data.

**IMPORTANCE**

The main purpose is to create prediction algorithms that can precisely calculate used device prices for new, untested data. When a model learns irrelevant patterns or noise from training data, it becomes overfitted and performs negatively on the data. A 70-30 split reduces the chance of overfitting and increases the models' capacity for generalization by providing an adequate quantity of data for training. Overall, this partitioning facilitates the development of reliable predictive models that can be integrated into the "Predict & Win" gaming platform, offering users accurate estimations of used device prices.

# **TABLET AND PHONE DATASETS**

Screen sizes in the dataset range from 5 cm to 30 cm, with bigger sizes possibly referring to tablets instead of smartphones. It is essential to distinguish between these two device groups to guarantee precision and clarity in our prediction models. This classification will enable more accurate predictive modeling based on the unique features of each type of device.

So, the dataset division will be seen in further steps.

# **GOAL 1: PREDICTING THE PRICE OF A USED DEVICE**

The platform enables customers to estimate the resale worth of the devices with accuracy by projecting the price of a used mobile. This provides insightful information about the potential value of their devices, which improves the user experience overall. Since "Predict & Win" revolves around estimating device costs, precise forecasts add to the gamification element of the program. As part of the game, players guess prices and aim to make the most accurate predictions and win rewards.

Regression models are used for this purpose and to build the regression models, it is important to start by taking the entire data into a new dataset and making necessary transformations before building the models so that the original data is not disturbed and is intact. This involves the following steps:

## **DATA PREPARATION FOR REGRESSION MODELS**

To build the regression models, it is important to start by taking the entire data into a new dataset and making necessary transformations before building the models so that the original data is not disturbed and is intact. This involves the following steps:

**Step 1:** **Factoring variables**

Transforming categorical variables into numerical representations, we allow simpler interpretation and analysis within our regression models, enhancing the effectiveness of the Predict & Win gaming platform.

**Operating System (OS):**

|  |  |
| --- | --- |
| Original | Transformed |
| Android | 1 |
| iOS | 2 |
| Windows | 3 |
| Others | 4 |

Table 2: Transforming Os

**Network Type:**

|  |  |
| --- | --- |
| Original | Transformed |
| 4g | 1 |
| 5g | 2 |
| Other | 3 |

Table 3: Transforming Network Type

**Brand:**

|  |  |
| --- | --- |
| Original | Transformed |
| Honor | 1 |
| Others | 2 |
| HTC | 3 |
| Huawei | 4 |
| Lava | 5 |
| Lenovo | 6 |
| LG | 7 |
| Meizu | 8 |
| Micromax | 9 |
| Motorola | 10 |
| Nokia | 11 |
| OnePlus | 12 |
| Oppo | 13 |
| Realme | 14 |
| Samsung | 15 |
| Vivo | 16 |
| Xiaomi | 17 |
| ZTE | 18 |
| Apple | 19 |
| Asus | 20 |
| Coolpad | 21 |
| Acer | 22 |
| Alcatel | 23 |
| Blackberry | 24 |
| Celkon | 25 |
| Gionee | 26 |
| Google | 27 |
| Karbonn | 28 |
| Microsoft | 29 |
| Panasonic | 30 |
| Sony | 31 |
| Spice | 32 |
| Xolo | 33 |

Table 4: Transforming Brand

**Step 2: Dividing the Dataset into Phone and Tablets**

Dividing the dataset based on screen sizes into two categories, namely phones and tablets, is a crucial step in enhancing the predictive capabilities of Predict & Win. An average phone size might be up to 18 cm and beyond that can be considered as a tablet devices.

**Step 3: Partitioning the Datasets.**

|  |  |  |
| --- | --- | --- |
|  | Phone | Tablet |
| Train | 70%- 2225 records | 70%- 189 records |
| Test | 30%- 950 records | 30%- 80 records |

Table 5: Tablet and Phone Partitioning

## **REGRESSION MODELS**

Regression models are essential to the "Predict and Win" platform because they help users make decisions, drive gameplay factors, and increase user engagement in addition to giving consumers accurate pricing forecasts.

### **LINEAR REGRESSION MODEL**

Understanding the linear relationship between predictors and the target variable can be done with the help of linear regression. It offers understandable coefficients that show the intensity and direction of each predictor's association with the target. The model is built on training data and its performance is evaluated on test data.

### **STEPWISE REGRESSION**

To create regression models that are easier to understand, and more straightforward, stepwise regression helps in the selection of the most relevant predictors. Here the backward method is used. Choosing only those predictors that have a meaningful impact on predicting the target variable, avoids overfitting and enhances model generalization.

After the linear regression model is built, it is useful to evaluate the model's performance and consider the significant variables that can be compared with the other multiple linear models.

### **REGRESSION TREE**

Regression trees provide insights into the decision-making process and are also simple to understand and visualize. Pruning the regression tree using the best complexity parameter (cp) will give better results. Figure 24 and Figure 25 show the tree plots.

**Regression Tree - Tablet Dataset:**

Pruned Tree with all predictors: Tree with significant predictors:

|  |  |
| --- | --- |
| A diagram of a graph  Description automatically generated  Figure 24: Tablet- Regression Trees | A diagram of a graph  Description automatically generated |

**Regression Tree - Phone Dataset:**

Pruned Tree with all predictors: Tree with significant predictors:

|  |  |
| --- | --- |
| A diagram of a computer generated image  Description automatically generated with medium confidence  Figure 25: Phone- Regression Trees | A diagram of a graph  Description automatically generated with medium confidence |

#### **REGRESSION Models for Tablet Dataset**

|  |  |
| --- | --- |
| Models | Evaluation MAE |
| Linear Regression | 0.1879 |
| Linear Regression with significant predictors | 0.1810 |
| Stepwise Regression | 0.1860 |
| Regression Tree | 0.2149 |
| Regression Tree with significant predictors | 0.2119 |

Table 6: Tablet- Regression Models

#### **REGRESSION Models for PHONE Dataset**

|  |  |
| --- | --- |
| Models | Evaluation MAE |
| Linear Regression | 0.1781 |
| Linear Regression with significant predictors | 0.1777 |
| Stepwise Regression | 0.1777 |
| Regression Tree | 0.2278 |
| Regression Tree with significant predictors | 0.2232 |

Table 7: Phone- Regression Models

## **REGRESSION MODEL SELECTION**

**Tablet Dataset:** Among the models evaluated, **Linear Regression with Significant Predictors** exhibits the lowest MAE (0.1810) as seen in Table 6. This indicates that this model provides the most accurate predictions for used tablet device prices.

**Phone Dataset:** Both Linear Regression with Significant Predictors and Stepwise Regression show identical MAE values (0.1777) as seen in Table 7. Given this, selecting **Linear Regression with Significant Predictors.**

By selecting the most suitable regression model for each dataset, Predict & Win ensures more precise pricing estimations for both used tablets and phones. Efficient model selection also allows to allocation of resources effectively, focusing on the development and deployment of the most effective regression models. This optimization maximizes the platform's predictive capabilities while minimizing operational costs.

**Classifying the predicted price:**

After selecting the best regression models for phones and tablets, we can create a new column to show the predicted price categories high or low for easy understanding. Adding a new column with predicted price categories enhances the user experience within the platform. Users can quickly grasp the estimated value of their devices and make informed decisions regarding their participation in challenges or purchases on the platform.

We can also build classification models to predict the used price categories. As part of the **project's requirements to build classification models** for predicting the price categories of used devices, the following steps can be undertaken to achieve the classification goal.

# **GOAL 2: CLASSIFYING USED DEVICE PRICE CATEGORIES**

Categorizing predicted prices provides valuable market insights for both users and platform administrators. By analyzing the distribution of predicted price categories over time, administrators can identify trends, patterns, and changes in the used device market. This information can be used to improve pricing algorithms, update prediction models, and enhance the overall effectiveness of the platform. Classification also allows users to earn rewards or points based on the accuracy of their predicted price range category.

## **DATA PREPARATION FOR CLASSIFICATION MODELS**

To build the classification models, it is important to start by taking the entire data into a new dataset and making necessary transformations before building the models so that the original data is not disturbed and is intact. This involves the following steps:

**Step 1: Create a new column for the price category.**

Categorize the used price into two categories: "low" and "high". The threshold for categorization is based on the normalized used price, with values below 4.0 categorized as "low" and the rest as "high". This step is crucial as it defines the target variable for the classification models.

**Step 2: Removing Unnecessary Columns**

Variables like normalized used and new prices are not needed for classification so can be removed from the dataset. This step ensures that only relevant features are retained for model training, reducing noise, and improving model performance.

**Step 3: Tablet And Phone Division**

The dataset is divided into two subsets based on the screen size of the devices: tablets and phones. This division allows for separate modeling of tablet and phone devices, considering their characteristics and user preferences.

**Step 4: Partitioning The Datasets.**

|  |  |  |
| --- | --- | --- |
|  | Phone | Tablet |
| Train | 70%- 2224 records | 70%- 189 records |
| Test | 30%- 951 records | 30%- 80 records |

Table 8: Phone Tablet Classification Partitioning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Used price category in Tablet | Train | | Test | |
| High | Low | High | Low |
| 186 | 3 | 79 | 1 |

Table 9: Tablet-Train And Test Records For Target

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Used price category in Phone | Train | | Test | |
| High | Low | High | Low |
| 1666 | 558 | 713 | 238 |

Table 10: Phone- Train And Test Records For Target

From Table 9 and Table 10, it's evident that there's a significant class imbalance between the "High" and "Low" categories in both the training and test datasets. This class imbalance can pose challenges for classification models because they may become biased towards predicting the majority class which is “High” due to its higher frequency. As a result, the model's performance metrics may not accurately reflect its ability to generalize to new data, especially for the minority class.

Despite the observed imbalance in the distribution of the used price categories, **we will proceed with building** the classification models using these partitions.

## **CLASSIFICATION MODELS**

### **LOGISTIC REGRESSION**

The glm() model is well suited for classification problems that give probabilities of belonging to a particular class. This is critical in a marketing environment where knowing the likelihood of a response is useful. The implementation and interpretation of logistic regression are made easier by its computing efficiency and simplicity.

### **DECISION TREE**

Decision trees can be used to identify the most significant features (e.g., brand, screen size, RAM) that contribute to the classification of devices into low or high-price categories. By analyzing these decision paths, users can gain insights into the factors affecting device pricing, enhancing their understanding of the market dynamics. Figure 26 shows the pruned tree for Phone Dataset.

A diagram of a cell phone

Description automatically generated

Figure 26: Phone- Classification Tree

### **K-NEAREST NEIGHBORS**

KNN can be utilized to classify devices into price categories based on their similarity to other devices in the dataset. By considering the characteristics of neighboring devices, KNN provides a flexible and adaptable approach to price category prediction, that shows varying market trends and user preferences.

#### **Classification Models for Tablet Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Accuracy | Sensitivity | Specificity |
| Logistic Regression | 5% | 3.8% | 1 |
| KNN | 98.75% | 100% | 0 |

Table 11: Tablet- Classification Models

#### **Classification Models for Phone Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Accuracy | Sensitivity | Specificity |
| Logistic Regression | 10% | 6% | 23% |
| Classification Tree | 89.17% | 95.93% | 68.91% |
| KNN | 88.01% | 93.41% | 71.85% |

Table 12: Phone- Classification Models

Tables 11 and 12 show the performance of the models for both tablet and phone datasets.

**These models are included as part of the project requirements but will not be utilized. Instead, the focus will be on categorizing the predicted used prices for better understanding and user engagement.**

# **NEW AND USED PRICE COMPARISON FOR PHONE AND TABLET DATASETS**

The gaming platform acquires useful insights into the distribution of pricing trends by calculating the percentage of new and used prices within each dataset. Users can discover a great deal about the current pricing dynamics in the tablet and phone marketplaces from these percentages.

The percentage of used devices in both datasets as shown in Figure 27, suggests that customers may be price-sensitive and might be willing to purchase used devices to save money. Therefore, offering affordable gaming options and promotions for used devices could attract more customers to the platform.

**A comparison of a pie chart

Description automatically generated**

Figure 27: Price Comparison.

# **CONCLUSION**

In conclusion, the "Predict & Win" project has successfully developed predictive models for estimating used device prices, enhancing user engagement and market insights within the gaming platform. By utilizing these models, the platform increases user satisfaction and engagement by giving consumers useful market data in addition to precise pricing estimates. The platform will continue to be a reliable and entertaining place to anticipate and win prizes based on used device values if models and user engagement tactics are continuously improved.

# **BUSINESS RECOMMENDATIONS**

The platform can focus on several avenues to further enhance the platform's predictive capabilities and user engagement. This may include expanding the dataset to include a wider range of device attributes and market factors for more comprehensive analysis and implementing user feedback mechanisms to continuously improve model accuracy and relevance to user needs. Additionally, exploring innovative gamification features and partnerships with device manufacturers or retailers could further enrich the user experience and drive platform growth.

# **EXECUTIVE SUMMARY**

**NAME**: Sreeja Reddy Singidi

**DATE**: 05-08-2024

**OPPORTUNITIES:**

The "Predict & Win" project offers an exciting chance to transform the market for used devices. Through the integration of gamification, e-commerce, and predictive analytics, the project aims to provide consumers with an interesting and unique platform for assessing the worth of their devices. Moreover, the project's focus on predictive analytics provides an opportunity to deliver valuable insights into market trends and pricing dynamics. Overall, the "Predict & Win" project represents a promising opportunity to create a dynamic and innovative platform that adds significant value to the used device market.

**SOLUTIONS:**

To achieve accurate price estimations for used smartphones and tablets on the Predict & Win platform, an approach is proposed. This involves implementing predictive analysis techniques such as regression modeling and classification algorithms to predict device prices based on various features. Additionally, by focusing on user engagement strategies and continuous improvement, the platform can establish itself as a leader in the device prediction space, capturing a significant share of the market and driving revenue growth.

PROJECT 3: MORTGAGE PAYBACK BEHAVIOR ANALYTICS

# **INTRODUCTION**

The dynamics of mortgage payback play a crucial role in the complex world of real estate financing, impacting loan companies' stability and profitability as well as the overall financial system. It's like solving a puzzle about how borrowers manage their mortgage payments. This project is all about figuring out patterns in how people repays their mortgages.

In the world of real estate finance, knowing how people pay back their mortgages is important. Mortgage payback analysis helps us understand how homeowners repay the money they borrowed to buy their houses. By looking at past data about when people got their mortgages and how they paid them back, we can learn a lot about how mortgages work. This analysis helps lenders and borrowers make better decisions about loans and housing. Lenders can adjust their risk management procedures and lending strategies by having a better understanding of the patterns observed.

# **PROJECT OVERVIEW**

Figures 1 and 2 give an overview of the project.

A screenshot of a diagram

Description automatically generated

Figure 1: Overview 1

*A diagram of a company

Description automatically generated*

Figure 2: Overview 2

# **BUSINESS UNDERSTANDING**

Understanding the dynamics of mortgage payback is crucial for various stakeholders in the real estate finance industry, including lenders, borrowers, and policymakers. For loan companies, staying stable and making money mostly depends on how well borrowers pay back their mortgages. If they understand how borrowers usually repay their loans, they can predict when they'll get money back, handle their cash well, and use their resources wisely. By spotting trends in how people pay back mortgages, they can plan better to make as much profit as possible while keeping financially safe.

Agencies keep a close eye on the mortgage business to make sure people are treated fairly and the financial system stays strong. When lenders look at how mortgages are paid back, they can show they're following the rules set by the government. This includes being fair to borrowers, managing risks properly, and being transparent about the terms of the loans. By understanding repayment patterns, lenders can also get ahead of any new rules coming up and avoid getting into trouble with the law.

The status of the loan tells what's happening with each loan in the dataset- whether it's been paid back, not paid back, or still ongoing. When we look at this, we can see if a loan has had problems in the past, like not being paid back on time. If a loan has had problems before, it might mean it's risky to give out another loan and can be considered a **red flag**.

# **BUSINESS GOAL**

The main business goal is to look at how people paid back loans before to see if they might have trouble paying back a new loan. This helps the lending company avoid losing money if borrowers can't pay back their loans on time. To precisely determine **credit risk**, they make use of historical data and predictive analytics. Lenders can reduce risks and decrease losses by using the financial system to identify high-risk applicants and likely default scenarios early. If a system indicates a default status it shows that loans have not been paid. When the system spots these defaults early, it can act like rejecting risky loan applications or changing the loan terms to lower the chance of future defaults. This way, it helps the lender avoid losing money and keeps their loan portfolio in good shape.

It is more important to use information about how people pay back their loans to make more money. This could mean finding people who are likely to pay back their loans on time and offering them special deals. It might also involve changing how much interest is charged based on how reliable borrowers have been in the past, or making the rules for getting a loan stricter to attract people who are more likely to pay it back.

**Two Approaches:**

* Classifying a customer as whether they will pay off the loan or default.
* Classifying a new customer into payoff or default based on the historic data.

# **ANALYTICAL APPROACH**

Before diving into data preprocessing, thorough analysis of the dataset is very important, and Excel can be most helpful in this case.

**Excel analysis:**

After analysis, it appears that the "status\_time" variable contains data on the default and payment times and 0 denotes that the debt has not been paid off or is in default. Finding loans that are at risk of default or that have already defaulted is the business's top priority because these situations call for quick attention and possible action to reduce losses. We can determine whether a record falls into the category of payoff or default by analyzing these 0 records.

To ensure the output generated can be reliable, we should seperate customers whose last observation is 0, which means it is non-default and non-payoff into a new dataset.

Combining details about the type of properties linked with mortgage loans into one category makes the data clearer. It helps us to understand the various property types easily.

Typically, credit scores range from 300 to 850, with 300 being the lowest score possible. However, in this dataset, the lowest acceptable credit score is 400, which is different from the usual range.

Moreover, finding 0 values in the interest rate column needs more investigation to decide how to deal with them in the next stages because no loan is given without the interest rate.

**Classification Algorithm Development:**

Following thorough analysis and data preprocessing, the next step involves building classification algorithms. By assessing the probability of default or payoff, these algorithms are essential in evaluating the financial ability of loan applicants. Lenders can improve risk management and improve the loan process by using these forecasts to guide their decisions on whether to approve or deny loan applications.

# **ATTRIBUTE DEFINITION**

1. id: In the dataset, every mortgage borrower is individually identified by the "id" field. The dataset makes it possible to follow and analyze the performance and behavior of individual borrowers over time by giving each borrower a unique ID. This has 50,000 unique observations.
2. time: It represents the period at which the observation was recorded for each mortgage borrower. Time attribute has 60 time periods or observations, capturing data over some time for each loan.
3. orig\_time: The beginning of a loan's lifespan is shown by its origination time. It denotes the specific time when the borrower enters into a loan agreement with the lender, by signing the mortgage documents and officially initiating the loan process. Some loans originated before the start of the period are indicated by negative values.
4. first\_time: This indicates the initial observation time that has been recorded. It refers to the first time that information about the loan—including its balance, interest rate, and other details—was recorded or noticed. It shows starting from the earliest observed period (1) to the latest observed period (60).
5. mat\_time: The time at which the mortgage loans mature is the "mat\_time" attribute. It indicates the point in the loan term that the borrower has taken to pay back the entire loan balance plus any interest that has accumulated. The customer can also pay the entire amount before the maturity time.
6. balance\_time: This shows the remaining balance of the mortgage loan at a specific time point. It reflects the amount of money still owed by the borrower to the lender at each observation time.
7. LTV\_time: LTV stands for Loan-to-Value ratio. It is a financial term used by lenders to show the ratio of a loan to the appraised value of the property securing the loan which means- it measures the percentage of the property's appraised value that is being financed by the loan.
8. interest\_rate\_time: Shows the interest rate on the mortgage loan at the observation time. This represents the percentage of the loan amount that the borrower must pay to the lender as interest for borrowing the money.
9. hpi\_index: The House Price Index is a metric used to monitor changes in the cost of residential property over time. The base year is 10. It offers insights into the general patterns and movements in property values within a certain region. If the HPI for a specific area increases by 5% in a given period, it indicates that, on average, house prices in that area have increased by 5% relative to the base year. In the same way, a decrease in the HPI suggests a decline in property values.
10. gdp\_time: It is the Gross Domestic Product (GDP) at the observation time. This allows analysts to evaluate the state of the economy during the period. GDP fluctuations throughout time may be a sign of changes in economic activity or recession. These changes may affect the mortgages.
11. uer\_time: Represents the unemployment rate at the observed time. This is the percentage of individuals who are unemployed and are actively looking for work among the total labor force within the given timeframe.
12. REtype\_CO\_orig\_time: Real estate type- condominium- at loan origination. This categorical attribute specifies the type of real estate associated with the mortgage loan at the time of origination, indicating if it's a condominium. The condominium is nothing but a condo which is a type of housing (apartment).
13. REtype\_PU\_orig\_time: Real estate type - planned urban development - at loan origination. Similar to the previous attribute, this categorical attribute denotes the type of real estate associated with the loan at origination, indicating if it's a planned urban development.
14. REtype\_SF\_orig\_time: Real estate type - single family - at loan origination. This categorical attribute indicates the type of real estate associated with the loan at origination, indicating if it's a single-family property.
15. investor\_orig\_time: This indicates whether the borrower is an investor or not. If the borrower is an investor, the value would be 1, and if not, it would be 0.

The borrower is an individual who invests in loans rather than borrowing money for personal or business use. Investors provide funds to borrowers in exchange for earning interest or returns on their investments.

1. balance\_orig\_time: Original balance amount of the loan. Like the balance\_time, this attribute sh ows the initial balance of the mortgage loan at the time of origination.
2. FICO\_orig\_time: This refers to the FICO credit score assigned to an individual at the time when their mortgage loan originated. It is crucial to figure out how risky the borrower is and how likely it is that they will make timely loan repayments. To determine the borrower's credit risk and to make well-informed judgments on loan approval, terms, and interest rates, lenders utilize FICO scores. Higher FICO scores are often associated with lesser credit risk and might lead to better loan terms for the borrower.
3. LTV\_orig\_time: Loan-to-value (LTV) ratio at loan origination. This denotes the loan-to-value ratio at the time of loan origination, giving insights into the initial financing structure of the loan.
4. Interest\_Rate\_orig\_time: Interest rate at loan origination. Like the interest\_rate\_time, this shows the annual interest rate charged on the mortgage loan at the time of origination.
5. hpi\_orig\_time: House price index at loan origination. Like the hpi\_time, this attribute gives information about the house price index at the time of loan origination, providing information on the property market conditions at the loan's inception.
6. default\_time: This binary attribute indicates whether the mortgage loan has defaulted (1) or not (0) at the specific observation time. When a loan defaults, it means that the borrower has not made the required payments as scheduled, which is going against the contract with the lender.
7. payoff\_time: Payoff time refers to the point in time when a borrower completes the repayment of their mortgage loan in full. When the payoff time is 1, it shows that the mortgage has been completely paid off by the borrower. Same way, if the payoff time is 0, it means that the loan has not been fully repaid.
8. status\_time: Provides information about the loan's status at a specific observation time. This attribute helps in understanding the current condition of loans within the dataset. It is a key attribute for assessing the financial health of the loan portfolio and making decisions about risk management and servicing.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Type |
| id | Identifier for each mortgage borrower. | Numeric |
| time | Time period for each record. | Numeric |
| orig\_time | Original time period when the mortgage loan was originated | Numeric |
| first\_time | First observation time period for the mortgage borrower | Numeric |
| mat\_time | Maturity time period for the mortgage loan. | Numeric |
| balance\_time | Balance amount remaining on the mortgage loan at the observation time | Numeric |
| LTV\_time | Loan-to-Value (LTV) ratio at the observation time | Numeric |
| interest\_rate\_time | Interest rate on the mortgage loan at the observation time | Numeric |
| hpi\_time | House Price Index (HPI) at the observation time | Numeric |
| gdp\_time | Gross Domestic Product (GDP) at the observation time | Numeric |
| uer\_time | Unemployment rate at the observation time | Numeric |
| REtype\_CO\_orig\_time | Real estate type (condo) at the original loan origination time | Categorical |
| REtype\_PU\_orig\_time | Real estate type (planned urban development) at the original loan origination time | Categorical |
| REtype\_SF\_orig\_time | Real estate type (Single Family Home) at the original loan origination time | Categorical |
| investor\_orig\_time | Investor being a borrower is given by 1 otherwise 0. | Categorical |
| balance\_orig\_time | The original balance amount of the mortgage loan at origination time | Numeric |
| FICO\_orig\_time | FICO credit score at origination time | Numeric |
| LTV\_orig\_time | Loan-to-Value (LTV) ratio at origination time | Numeric |
| Interest\_Rate\_orig\_time | Interest rate on the mortgage loan at origination time | Numeric |
| hpi\_orig\_time | House Price Index (HPI) at origination time | Numeric |
| default\_time | Default status (0 for non-default, 1 for default) | Categorical |
| payoff\_time | Payoff status (0 for non-payoff, 1 for payoff) | Categorical |
| status\_time | Status indicator (0 for non-default/non-payoff, 1 for default, 2 for payoff). | Categorical |

Table 1: Attributes

# **DATA UNDERSTANDING**

Before proceeding with data preprocessing, it's crucial to thoroughly understand the dataset and make necessary adjustments, such as handling missing and zero values. Handling missing and zero values is particularly important as it ensures that analysis is based on complete and accurate information. This helps to better understand the behavior of each borrower by having a clear view of their data. These steps are **performed first** because based on all the records of each borrower we will know how they are behaving and **take appropriate measures** before removing the records.

## **NUMBER OF RECORDS**

This is a huge real-time dataset that consists of 622490 observations and 23 variables that are the characteristics of the mortgage.

## **CHECKING FOR MISSING VALUES**

We see from Figure 3, that there are 270 missing values and all of them are from LTV time.

A blue and black graph

Description automatically generated with medium confidence

Figure 3: Missing values

It has been noticed that these 270 records are from 18 different borrowers. Table 2, shows how many missing records are observed for each borrower ID.

|  |  |
| --- | --- |
| Borrower ID | Missing records |
| 39722 | 36 |
| 39723 | 36 |
| 39724 | 1 |
| 39725 | 36 |
| 39726 | 4 |
| 39727 | 10 |
| 39728 | 3 |
| 39729 | 2 |
| 39730 | 36 |
| 39731 | 19 |
| 39732 | 5 |
| 39733 | 36 |
| 39734 | 3 |
| 39735 | 17 |
| 39736 | 5 |
| 39737 | 8 |
| 39738 | 4 |
| 49658 | 9 |
| Total | **270** |

Table 2: Missing Values For Each Borrower

Grouping by ID for imputation is impossible due to the pattern seen in the dataset, where most IDs had records with missing values in the "LTV\_time" column only. As of now, eliminating all IDs that have missing data.

## **CHECKING FOR 0 VALUES**

It was found that multiple columns in the dataset had 0 values when the dataset was examined for 0 values. It's crucial to remember that some of these characteristics- like real estate types, investor origin times, payoff and default times, and status times—are binary variables. When it comes to binary variables, one category (like non-default, not paid off, or a particular kind of real estate) is usually represented by a value of 0, and the other (like default, paid off, or a different kind of real estate) by a value of 1. As a result, the 0 values in these columns reflect the absence of a category rather than necessarily indicating missing or incorrect data. Because these 0 values contain important information, they shouldn't be regarded as true missing data.

A screen shot of a computer

Description automatically generated

Code Snippet 1: Zero Values In Each Attribute.

**Origin time:**

The 0 values in "orig\_time" and the consistent value of 102.24 for "hpi\_orig\_time" from all these 0 values could mean that these loans originated during a time when the house price index remained relatively stable. It might be a default placeholder value applied to loans that were made during periods or under conditions.

These 0’s can also mean that the loans originated before the start of the observation period. This happens because some loans are already in existence when the data collection begins, and their exact origination time might not be available due to privacy or security concerns. So, instead of having specific origination timestamps, these loans are represented with 0 values in the dataset.

Negative values indicate that the loan was initiated sometime before the beginning of the data collection period.

**Balance time:**

When a borrower's balance on a mortgage loan is 0, it shows that they have paid back the entire loan amount. To confirm this, checked the status time which shows that that loan status is either paid off or still going on for all these records.

**LTV time:**

When the loan's outstanding balance is zero, it indicates that the borrower has paid back the entire loan amount. In these situations, as there is no longer a loan balance owed about the property's value, the LTV (Loan-to-Value) ratio would anyway be zero.

**Interest rate time and Interest rate origin time:**

Interest rate is shown in percentage terms that the lender is charging to borrowers. I’ve noticed that all these are default types of loans. These need to be handled because according to my knowledge, every loan will have at least a minimum interest rate.

Handling these 0’s can be done by imputation. Correctly handling these 0 values plays a major role in achieving the business goal.

In this case, it is handled by replacing 0 values with the median interest rate for each borrower ID but noticed that 2 IDs still have 0 interest rates. Those borrower ID records can be eliminated.

To handle 0 values in Interest rate origin time, the dataset and be grouped by borrower ID, and the median interest rate for each ID can be calculated based on the interest\_rate\_time column.

The median interest rate for each borrower ID can be then used to replace 0 values in the Interest\_Rate\_orig\_time column, ensuring that each borrower's interest rate is represented accurately.

# **DATA PREPROCESSING**

## **OPTIMIZING DATASET FOR MODEL EFFICIENCY**

In line with the business goal of accurately classifying customers as defaulters or payoffs, it's essential to simplify the dataset to enhance model efficiency. This involves examining the usefulness of records containing multiple timestamps leading up to deployment.

There are few **borrowers** who have **only one observation**, and one observation is not sufficient to make meaningful conclusions about the behavior of a borrower. So, all the borrowers with just one observation can be **excluded**.

In the dataset, each borrower has multiple records corresponding to different time stamps, capturing various attributes such as loan balance, LTV ratio, interest rate, and more. However, for analysis and modeling purposes, it's often more convenient to have consolidated data where all relevant information is captured into a single observation of a borrower.

The **goal** is to create a more **condensed** and structured dataset that retains all essential information while simplifying the analysis process.

This process works like:

The **first step** is to identify a unique identifier for each borrower, that is ID. This identifier will be used to group the data during the aggregation process.

In the **second step**, we group the dataset by the ID, ensuring that all records for each borrower are grouped together.

**Next**, within each group, we aggregate or summarize the information across different time stamps. This could involve statistics, or simply selecting the most recent value.

**Finally**, a new dataset will be created where each row represents a unique borrower, with all relevant information aggregated into columns. This condensed dataset is more suitable for analysis and modeling tasks.

All this will be performed by using **sqldf** in R so that the process does not take much time and is easy to understand.

## **INITIAL TRANSFORMATION**

### **TIME ATTRIBUTE TO OBSERVATION COUNT**

The time attribute consists of all the records of each borrower where he made a payment for each time stamp. This is not necessary, so we can consider how many time stamps or transactions are associated with each borrower. This calculation is helpful because it provides a quick overview of the activity of each borrower.

### **ORIGINATION TIME**

This is the time stamp for origination which will be same for borrower. So, considering the last time stamp’s corresponding value will be sufficient.

### **FIRST TIME**

This is the time stamp of the first observation. Like the origination time this can also consider the last time stamp’s value.

**Maturity attribute** is not considered in this new dataset because this dataset deals with 30 years of maturity only. As it is evident that all the borrowers have the same maturity time, we can exclude this column.

### **HANDLING BALANCE TIME COLUMN**

Consider the **last time stamp** observation for balance time.

Also add a **new column** to show how many **times** a **borrower did not make any payment** and balance remained same around these time stamps.

### **LTV RATIO**

The LTV ratio is a measure of the loan amount compared to the appraised value of the property securing the loan. It is considered to average the LTV for each borrower because watching how the average LTV ratio changes over time can tell us a lot about how risky a borrower's loans are becoming. For instance, if the average LTV ratio is going up, it could mean borrowers are borrowing more money compared to the value of their properties.

### **INTEREST RATE**

The average is calculated the same as LTV ratio.

### **FIXED INTEREST RATE**

A **new column** can be added based on Interest rate. The **fixed interest rate** indicates whether the interest rate remains constant or changes over time for each borrower's loan. The value 1 suggests that the interest rate remains fixed throughout the loan term, while a value of 0 indicates that the interest rate is variable and may change over time. If the interest rate remains fixed, borrowers may repay, leading to payoffs. If the interest rate is 0, fluctuations in the rate could impact borrower’s ability to make timely payments, resulting in defaults.

### **HOUSE PRICE INDEX, UNEMPLOYEMNT, GDP**

All these 3 are the economic indicators which will be same for each time stamp. Here the latest observation can be considered.

### **REAL ESTATE TYPES**

Based on the original type of real estate, each ID is categorized into four types: Single Family as these have majority records in the dataset as 1, Planned Urban Development as 2, Condo as 3 and 4 means other type of real estate. All these are merged into a single attribute called real estate type.

### **INVESTOR TYPE**

The investor type for each ID at the latest time stamp shows whether the borrower is an investor or not. In the future, we can analyze these investor records separately.

### **ORIGINATION TIME COLUMNS HANDLING**

Various origination time-related columns (such as balance time, FICO, LTV, interest rate and hpi index origination) for each ID at the latest time stamp are extracted.

### **STATUS TIME**

The status time for each ID at the latest time stamp is obtained.

Default time and pay off time are removed from the data because they contain the same information as in status time.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Type |
| id | Identifier for each mortgage borrower. | Numeric |
| observation\_count | How many time stamps or transactions are associated with each borrower. | Numeric |
| origin\_time | Last original time period when the mortgage loan was originated | Numeric |
| first\_time | First observation time for the mortgage borrower | Numeric |
| zero\_diff\_count | How many times the borrower did not make any payment between any 2-time stamps. | Numeric |
| ltv\_ratio | Average of the LTV time for each borrower | Numeric |
| interest\_rate | Average Interest rate on the mortgage loan at the observation time | Numeric |
| fixed\_interest | The interest rate remains constant or changes over time for each borrower's loan. 1= fixed, 0= not fixed | Categorical |
| hpi\_time | House Price Index (HPI) at the observation time | Numeric |
| gdp\_time | Gross Domestic Product (GDP) at the observation time | Numeric |
| uer\_time | Unemployment rate at the observation time | Numeric |
| real\_estate\_type | 1 = single family, 2 = planned urban development, 3 = condo, 4 = Others | Categorical |
| investor\_orig\_time | Investor being a borrower is given by 1 otherwise 0. | Categorical |
| balance\_orig\_time | The original balance amount of the mortgage loan at origination time | Numeric |
| FICO\_orig\_time | FICO credit score at origination time | Numeric |
| LTV\_orig\_time | Loan-to-Value (LTV) ratio at origination time | Numeric |
| Interest\_Rate\_orig\_time | Interest rate on the mortgage loan at origination time | Numeric |
| hpi\_orig\_time | House Price Index (HPI) at origination time | Numeric |
| status\_time | Status indicator (0 for non-default/non-payoff, 1 for default, 2 for payoff). | Categorical |

Table 3: Final Attribute Table

# **SIGNIFICANCE OF ANALYZING ACTIVE LOANS**

Creating a **separate dataset** for loans with **status equal to 0** (indicating that the loan process is still ongoing) plays a major role as this dataset will be used for predicting whether the profile will pay or default on the loans.

By doing this, we can focus specifically on the loans that have not yet reached a conclusion, whether that be default or payoff. This allows us to analyze the characteristics and behaviors of these ongoing loans separately from those that have already defaulted or been paid off.

Once the models are trained and validated using historical data, they need to be evaluated on **unseen data** to assess their real-world performance. The ongoing loans dataset acts as this unseen data, allowing us to assess how well the models generalize to new instances and predict the outcomes of ongoing loans.

# **PREDICTOR ANALYSIS**

## **REAL ESTATE TYPES**

The graph from Figure 4, shows the distribution of real estate types and provides insights.

It is evident from Figure 4, that Single-family homes might dominate the local real estate market, leading to a higher proportion of loans secured by this property type. This might result from factors such as affordability or cultural preferences. Single-family homes might respond differently to economic fluctuations compared to other property types, showing their default and payoff rates.

Condos and urban developments have lower counts overall, indicating a smaller proportion of loans associated with these property types.

A graph of different colored squares

Description automatically generated

Figure 4: Real Estate Distribution.

## **ANALYZING ATTRIBUTE INFLUENCE ON DEFAULT AND PAYOFF**

The factors that influence mortgage default or payoff enable lenders to make more informed decisions about loan approval and interest rates. Lenders can adjust their lending criteria and pricing based on the risk profile associated with specific attributes, ultimately giving profits, and minimizing losses. This can also identify trends, patterns, and potential areas of concern.

### **INVESTORS STATUS**

Investor status allows lenders to make informed decisions about whether to approve loan applications from investors and non-investors. If loans initiated by investors have higher default rates, lenders may choose to apply stricter criteria or higher interest rates for such applicants.

From Figure 5, most borrowers, regardless of their loan outcome, are regular individuals rather than investors. The higher presence of investors among defaulters compared to payoffs could indicate potential differences in risk-taking behavior or investment strategies between investors and non-investors. Investors might be more willing to take on higher-risk loans, leading to a higher likelihood of default in **some** cases. We can provide **rewards to attract** borrowers who are investors.

A red and blue bars

Description automatically generated

Figure 5: Comparison of Investor’s Distribution

## **LOAN ORIGINATION ATTRIBUTES**

### **RELATION BETWEEN ORIGIN TIME (DEFAULT AND PAYOFF)**

The plot gives insights into how the status of loans has changed over time since origination. This understanding is valuable for assessing the performance of loans and identifying potential risk factors associated with origination time.

It is noticed that at the period of 23-25 months, there is a highest peak in the defaulter's status which shows that most loans are defaulted on in that period. Knowing when loans are most likely to go into default helps lenders use their money better, make specific plans, and change how loans work to lower risks and make more money overall.

A graph of a person with red squares

Description automatically generated with medium confidence

Figure 6: Analysis Of Origination Time.

### **FICO AMONG LOAN STATUS**

Comparing FICO scores at the time of loan origination between defaulters and payoffs allows lenders to assess the success of their credit policies. In Figure 7, it's evident that loans given to borrowers with higher FICO scores tend to have lower default rates. Additionally, looking at the distribution of FICO scores for payoffs, we observe that borrowers with higher credit scores are more likely to successfully pay off their loans. This suggests that prioritizing borrowers with better credit scores can be an effective strategy for reducing default rates and improving loan repayments. Such insights help lenders make informed decisions about borrower eligibility criteria and risk management strategies.

|  |  |
| --- | --- |
| A graph of a graph  Description automatically generated with medium confidence | A graph of a graph  Description automatically generated |

Figure 7: Distribution Of Credit Score

### **DISTRIBUTION OF LTV AND LTV ORIGIN**

The distribution of Loan-to-Value (LTV) ratios at a specific time and at the time of loan origination provides insights into the levels of borrowers. By analyzing these distributions, we can understand how the relationship between loan balances and property values changes over time.

A peak in the range of 50-100 suggests that a significant portion of borrowers have relatively high LTV ratios, indicating that they have borrowed a substantial amount compared to the value of their properties. This could imply higher risk for lenders, as borrowers with higher LTV ratios may be more vulnerable to default if property values decline or if they encounter financial difficulties.

A distribution concentrated below 100 suggests that many borrowers initially had lower LTV ratios, indicating that they had more equity in their properties at the time of loan origination. This could be less risky for lenders, as borrowers with lower initial LTV ratios may have a greater financial stake in their properties and may be less likely to default.

A graph of a bar graph

Description automatically generated with medium confidence

Figure 8: Ltv and Ltv Origin

## **RELATION BETWEEN GDP AND UNEMPLOYMENT RATE**

When the GDP of a country decreases, it often indicates a slowdown in economic activity. This slowdown may lead to reduced business investments, lower consumer spending, and decreased demand for goods and services. Consequently, businesses may lay off workers or reduce hiring, resulting in a higher unemployment rate.

When the GDP grows, it signifies an expansion in economic activity. This growth can stimulate business investments, boost consumer confidence, and increase demand for products and services. As businesses expand to meet this demand, they may hire more workers, leading to a decrease in the unemployment rate.

A graph with blue dots

Description automatically generated

Figure 9: Relation Between Gdp And Uer

## **FIXED INTEREST RATE WITH STATUS TIME:**

Loans with fixed interest rates tend to have a higher probability of being paid off compared to defaulting. This suggests that borrowers whose interest rate is fixed may have more stable financial circumstances contributing to successful loan repayment and a reduced no. of defaults.

A red and blue squares

Description automatically generated

Figure 10: Fixed Interest with status time

## **ECONOMIC INDICATORS FOR SELECTED TIME FRAME**

Economic indicators like GDP (the total value of goods and services in an economy), HPI (how housing prices change over time), and UER (the percentage of unemployed people), will be the **same** over each time stamp.

**Why does this matter for mortgages?**

When the economy is steady, people tend to feel more confident about borrowing money for things like house purchases. Also, lenders can better predict if borrowers will be able to pay back their loans on time from the unemployment rate. So, understanding when the economy is stable helps lenders make smarter decisions about who to give loans to and how much interest to charge. It's like having a clearer picture of how risky or safe it is to lend money at different times.

## **RELEVANCY**

Predictors like id and observation\_count have correlation close to 0 with most other predictors. This suggests that they may not be strongly correlated with other predictors and may not provide much additional information for modeling.

Also, investor\_orig\_time has low correlation with many other variables, indicating that it may not have strong linear relationships with other predictors, but low correlation may still be important predictors if their relationships with the status time are non-linear.

Now analyzing the status time, we notice that interest rate, investor origin time, balance origin, FICO and Interest rate origin, have weak correlations with status time.

For building the model all the predictors are considered.

A close-up of a graph

Description automatically generated

Figure 11: Correlation Plot

# **DIMENSION REDUCTION**

After all the preprocessing, again missing values and 0 values are checked. There are duplicate records of a few IDs which need to be handled.

When the same ID is listed more than once, it might mean there's a mistake in the dataset. Dealing with these duplicates shows we're making sure each piece of information is only counted once. This helps avoid having too much of the same data and keeps dataset accurate and reliable.

|  |  |
| --- | --- |
| ID | No. of duplicate records |
| 3014 | 63 |
| 36833 | 63 |
| 36910 | 63 |
| 37067 | 63 |
| 37130 | 63 |
| 37351 | 63 |
| 37528 | 63 |
| 37558 | 63 |
| 37580 | 63 |
| 37637 | 63 |
| 37792 | 63 |

Table 5: Duplicate Records

After removing these duplicate records, the no. of IDs in the data match with- after removing all the missing values from the original dataset.

# **DATA TRANSFORMATION**

The names of the attributes can be changed for better understanding. Changing to more descriptive and understandable terms enhances the clarity and interpretability of the data, making it easier for stakeholders to grasp the meaning of each variable.

|  |  |
| --- | --- |
| Original | Transformed |
| observation\_count | no\_of\_payments |
| first\_time | first\_payment\_time |
| zero\_diff\_count | missed\_payment\_times |
| hpi\_time | hpi\_index |
| investor\_orig\_time | investor |
| balance\_orig\_time | balance\_at\_origin |
| FICO\_orig\_time | credit\_score |
| LTV\_orig\_time | ltv\_origin |
| Interest\_Rate\_orig\_time | interest\_rate\_origin |
| hpi\_orig\_time | hpi\_origin |
| status\_time | status\_type |

Table 6: Data Transformation

# **DATA PARTITIONING METHODS**

**NECESSITY**

By dividing the dataset, we can train models on a portion of the data, assess their efficiency, and ensure the models operate well when applied to new, untested data. By preventing overfitting, optimizing model performance, and offering trustworthy smartphone pricing predictions, this method improves the effectiveness and reliability of the automated loan platform.

**DIFFERENT APPROACHES**

Data partitioning can be done in different methods:

**Train-test division**: The process of train-test partitioning involves splitting the dataset into two subsets: a training set for model training, and a testing set for model evaluation.

**Cross Validation**: It is a resampling strategy that divides the dataset into several subsets of training and validation sets iteratively to assess the performance of the models. Training and testing a model on several subsets of the data lowers the likelihood of overfitting and offers a more reliable evaluation of the model's performance, which shows how well the model will generalize to new data.

**APPROACH IN THIS PROJECT**

 The straightforward train-valid-test split approach will be used in this project, allocating **70%** of the dataset for training and **30%** for validation. This gives the models enough information to identify patterns and connections in the data and helps assess how well the trained models generalize to new data that is test data.

**IMPORTANCE**

It ensures that the classification models are robust and reliable when deployed in real-world scenarios, contributing to better decision-making and risk assessment.

# **GOAL: CLASSIFYING BORROWERS**

The primary objective is to identify individuals who are at risk of defaulting on their loans and may require additional assistance to ensure timely repayment. Those IDs might be targeted for loan repayment outreach. By accurately predicting which borrowers are more likely to default, this system can allocate resources and engage in targeted outreach efforts to reduce the risk of loan defaults. These IDs require additional support. Here we will be using **active records to test**, by which we can assess the effectiveness of the classification models in predicting the future outcomes of these borrowers.

## **MODEL SELECTIONS**

### **LOGISTIC REGRESSION**

The glm() model is well suited for classification problems that give probabilities of belonging to a particular class. The implementation and interpretation of logistic regression are made easier by its computing efficiency and simplicity. This is a simple yet powerful model for binary classification.

### **DECISION TREES**

Decision trees partition the feature space into regions and make predictions based on the majority class in each region. They are useful in this project for their simplicity and interpretability, making them easy to understand and explain to stakeholders.

### **K-NEAREST NEIGHBORS (KNN)**

KNN is a non-parametric algorithm that makes predictions based on the majority class of the k nearest neighbors. It might not be an appropriate approach for this project as KNN's performance decreases as the dimensionality of the data increases. This is known as the **"curse of dimensionality"**.

### **NAÏVE BAYES**

Naive Bayes classification is a probabilistic algorithm that is based on Bayes' Theorem, with the "naive" assumption that features are conditionally independent given the class label. This model works better with categorical features which means it might not be suitable in this case.

## **MODEL PERFORMANCE**

After building the models, their performance is evaluated. Sensitivity measures how well the models identify payoffs, while specificity shows their ability to recognize default loans correctly. It is important to assess sensitivity and specificity because they provide insights into how well the models perform in identifying payoffs and default loans, respectively. **KNN and Naïve Bayes** need **not be considered** for model selection as they might not be suitable for this dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Accuracy | Sensitivity | Specificity |
| Logistic Regression | 78.53% | 85.00% | 68.00% |
| Classification Tree | 76.10% | 83.80% | 63.55% |
| KNN | 75.50% | 80.54% | 67.28% |
| Naïve Bayes | 76.07% | 78.68% | 71.82% |

Table 7: Model Performance

According to the business goal it is **important** to identify the **default loans** so that the system can initiate communication with the defaulter to understand their circumstances and explore options for repayment or restructuring of the loan or may pursue legal action against the defaulter for repayment. The system can also engage **third-party debt collection agencies** to recover the outstanding debt on their behalf. These agencies specialize in contacting and negotiating with defaulters to secure repayment.

## **BEST MODEL**

Among the logistic regression, and classification tree models, the **logistic regression model** has the highest specificity of 68%. Specificity is crucial in this context because it represents the ability of the model to correctly identify default loans, which is essential for the business goal of taking appropriate actions.

Now considering **this model,** we need to predict on the **ongoing dataset** that is separated so that we will know who will pay and who will not pay in the future. By this we can take appropriate measures on defaulters.

# **PREDICTING THE STATUS ON ACTIVE LOANS**

Using this model, we will predict the loan repayment status of borrowers in an ongoing dataset. By doing so, it aims to distinguish between borrowers who are likely to fulfill their loan obligations and those who may default. This predictive insight will enable us to implement appropriate measures for managing default risks effectively.

The ongoing data consists of 8168 records where number of **defaulters predicted is 2310** and rest borrowers **5858, have paid** off according to the predictions. So, these defaulters need extra measures. Ultimately, by identifying potential defaulters beforehand, we can take proactive steps. This proactive approach helps to reduce potential losses and optimize loan portfolio management.

# **CONCLUSION**

This project highlights the critical importance of understanding mortgage repayment dynamics for lenders and borrowers in the real estate finance industry.

By analyzing historical repayment patterns and building predictive models, lenders can optimize their loan portfolio management strategies to minimize default risks and maximize profitability. By identifying individuals who are at risk of defaulting on their loans, lenders can allocate resources and handle the risk of loan defaults. This helps minimize financial losses, improve overall portfolio performance, and enhance customer satisfaction.

# **BUSINESS RECOMMENDATIONs**

Refine loan approval criteria based on insights from repayment patterns to attract reliable customers with lower default probabilities. Adjust interest rates, loan terms, and eligibility requirements to align with the risk profile of potential borrowers and optimize profitability.

# **EXECUTIVE SUMMARY**

**NAME:** Sreeja Reddy Singidi

**DATE:** 05-08-2024

**OPPORTUNITIES:**

The analysis of mortgage repayments presents great chances for banks to make things easier, lower risks, and make customers happier. By examining borrower information carefully, banks can make better decisions and spot problems early to prevent defaults. This helps banks follow rules, save money, and stay competitive in the loan market.

**SOLUTIONS:**

To achieve the goal, the system should focus on predicting if someone might not be able to pay back a loan so that special assistance will be provided to those borrowers. These classification models will look at information about borrowers and help the system make smart decisions about who to lend money to. Also, the **system** needs to keep an eye on how well it works and make **improvements** when needed to keep up with changes in the loan market.