North-Point Software Production Company

By Komal Singh

George Herbert Walker School of Business and Technology, Webster University

CSDA 6010: Data Analytics Practicum

Professor: Ali Ovlia

**Table of Contents**

**Introduction …………………………………………………………………………………………….3**

Business Problem ………………………………………………………………………………..3

Analytics Goal …………………………………………………………………………………...3

**Data Exploration and Pre-processing …………………………………………………………………4**

Introduction of New Features ……………………………………………………………………4

Attribute Definition ……………………………………………………………………………..4

Structure of the Data ………………………………………………………………………….....6

**Attribute Analysis ………………………………………………………………………………………7**

Scatter plot between Frequency and Spending variables. ……………………………………….7

Purchase and Spending status by Gender. ……………………………………………………….7

Box plot between spending and Re-update variables ……………………………………………8

Box plot between spending and marketplace US or not …………………………………………9

Bar plot between spending and all the sources ………………………………………………….10

Histogram for attribute spending ………………………………………………………………..10

Bar graph b/w address\_is\_res and Spending ……………………………………………………11

Bar graph b/w address\_is res and Purchase ……………………………………………………..12

**Predictor Analysis and Relevancy …………………………………………………………………….12**

Chi-Square Test for Predictor Correlation ………………………………………………………13

ANOVA for Predictor Relevancy ………………………………………………………………..13

**Dimension Reduction …………………………………………………………………………………..15**

**Feature Selection using Stepwise Regression with AIC ……………………………………………...15**

Feature Selection for purchase model …………………………………………………………...15

Feature Selection for Spending model …………………………………………………………..16

Interpretation of Coefficients ……………………………………………………………………17

**Data Sampling …………………………………………………………………………………………17**

**Classification Models for Purchase …………………………………………………………………...17**

Classification Model Selection ………………………………………………………………….21

**Regression Model for Spending ……………………………………………………………………….21**

**Model Evaluation on Holdout Data …………………………………………………………………..26**

**Gain Chart ……………………………………………………………………………………………...27**

**Lift Chart ……………………………………………………………………………………………….28**

**Gross Profit Estimation ………………………………………………………………………………..28**

**Business Recommendation …………………………………………………………………………….29**

**Executive Summary ……………………………………………………………………………………30**

**Introduction:**

North-Point is a software list firm that sells games and educational software. It started out as a software manufacturer and later added third-party titles to its offerings. It has recently put together a revised collection of items in a new Listing, which it is preparing to roll out in a mailing.

In addition to its own software titles, North-Point’s customer list is a key asset. To expand its customer base, it has recently joined a consortium of listing firms that specialize in computer hardware and software products. The consortium affords members the opportunity to mail lists of names drawn from a pooled list of customers. Members supply their own customers lists to the pool and can withdraw an equivalent number of names in each quarter. Members are allowed to do predictive modeling on the records in the pool so they can do a better job of selecting customer names from the pool.

The Mailing Experiment:

North-Point has supplied its customer list of 200,000 names to the pool, which totals over 5,000,000 names, so it is now entitled to pick 200,000 names for a mailing.

North-Point would like to select the name that has the best chance of performing well, so it conducts a test – it draws 20,000 names from the pool and does a test mailing of the new list.

This mailing yielded 1065 purchasers, a response rate of 0.053 or 5.3%. To optimize the performance of the machine learning models, it was decided to work with a stratified sample that contained an equal number of purchasers and non-purchasers. For ease of presentation, the data set in this project includes just 1000 purchasers and 1000 non-purchasers, an apparent response rate of 0.5 or 50%. Therefore, after using the dataset to predict who will be a purchaser, we must adjust the purchase rate back down by multiplying each “case’s probability of purchase” by 5.3/50 or 0.106.

**Business Problem:**

The primary business problem is to optimize the selection of customer names from a large pool to maximize the response rate in their mailing campaign for a new listing and to see if how many of the customers would be willing to spend.

**Analytics Goal:**

The goal is to use predictive modeling and data analysis to improve the effectiveness of North-Point's mailing campaign for their new software listing.

Identifying and selecting the most promising customer names from a large pool

using data analytics to predict which individuals are likely to make a purchase and the amount they will be spending on the purchase because it costs $2 to mail the booklets. And the company has 200,000 customers from the pool. If the company tries to send the mail to all 200,000 customers, it will cost them more.

Analytics Approach:

Building two machine learning models one with a target variable Purchase which would be to classify observations as purchase or no purchase and another model with a target variable as spending, for those cases that are classified as purchase and will predict the amount they will spend.

**Data exploration and preprocessing:**

Data Understanding:

The dataset contains both categorical and numerical variables, and the primary focus is on predicting customer purchasing behavior and spending patterns. The primary dataset we received had 25 variables and 2000 observations in it. two additional variables have been added to the dataset by combining existing variables or converting the categorical version of the numerical variables.

Observations:

* The dataset has been checked for missing values and no missing values were found in any of the columns.
* We have identified 64 records that do not correspond to any of the existing sources and do not have a web order.
* We found a single record where no purchase was made, but a spending of 1 occurred.

**Introduction of New Features:**

Two new columns have been added to the dataset, The binary nature of both columns enables straightforward filtering and categorization, helping in the identification of specific patterns or anomalies within our data.

Source\_unknown:

* The column 'source\_unknown' serves as a binary indicator, distinguishing records that do not align with any of the predefined sources or web orders. By aggregating 64 instances falling outside these categories, this binary column facilitates the identification and isolation of data points that may require special attention or further investigation.

Re\_update:

* The 'Re\_update' column shows us if something changed between the '1st\_update\_days\_ago' and 'last\_update\_days\_ago'. It's like looking at the dates when things were first updated and when they were last updated. If the difference between these dates is zero, it means nothing changed and there was only one update made to the customer (so 'Re\_update' is 0). But if the difference is not zero, it means the customer was updated more than once, and then 'Re\_update' is set to 1.

**Attribute Definition:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column No. | Variable Name | Description | Type | Code Description |
| 1 | sequence number | Row # |  |  |
| 2 | US | Is US address? | Binary | 1 = yes 0 = no |
| 3 to 17 | source\_\* | Source List for the record - 15 possibilities | Binary | 1 = yes 0 = no |
| 18 | Freq | Number of transactions in last year at source catalog | Numerical |  |
| 19 | last\_update\_days\_ago | How many days ago last update was made to customer record | Numerical |  |
| 20 | 1st\_update\_days\_ago | How many days ago first update to customer record was made | Numerical |  |
| 21 | Web order | Customer placed at least one order via web | Binary | 1 = yes 0 = no |
| 22 | Gender=male | Customer is male | Binary | 1 = yes 0 = no |
| 23 | Address\_is\_res | Address is residence | Binary | 1 = yes 0 = no |
| 24 | Purchase | Customer made purchase in test mailing | Binary | 1 = yes 0 = no |
| 25 | Spending | Amount in dollars customer spent in test mailing | Numerical |  |
| Additional Variables added to dataset | | | | |
| 26 | Re\_update | If the update to the customer was done once or twice | Binary | 1 = yes 0 = no |
| 27 | Source\_unknown | Records that don’t belong to any source or web order | Binary | 1 = yes 0 = no |

Numeric Data:

* Freq: This column represents the frequency or number of purchases made by the customer. It is likely a numeric variable that indicates the count of purchases.
* last\_update\_days\_ago: This column may indicate the number of days that have passed since the last update or interaction with the customer's data.
* 1st\_update\_days\_ago: It may represent the number of days that have passed since the first update or interaction with the customer's data.
* Spending: For those prospects who made a purchase (Purchase = 1), this variable represents the amount they spent. It is likely a numeric variable that indicates the monetary value of the purchase.

Categorical Data:

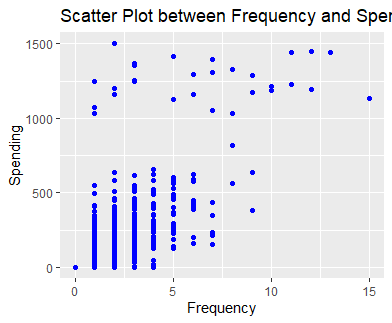
* US: This column indicates whether the customer is located in the United States or not. It is likely a binary variable, with 1 representing customers from the US and 0 representing customers from other countries.
* Source\_a, source\_c, source\_b, source\_d, source\_e, source\_m, source\_o, source\_h, source\_r, source\_s, source\_t, source\_u, source\_p, source\_x, source\_w: These columns likely represent different sources or channels through which customers were acquired. Each column may contain binary values (0 or 1) indicating whether a particular source was used to acquire the customer.
* Web order: This column may indicate whether the purchase was made through a web order or not. It could be a binary variable with 1 representing a web order and 0 representing other types of orders.
* Gender=male: This column may indicate the gender of the customer, with a binary value of 1 representing male and 0 representing female or other genders.
* Address\_is\_res: This column may indicate whether the customer's address is a residential address or not. It could be a binary variable with 1 representing a residential address and 0 representing a non-residential address.
* Purchase: This variable indicates whether a prospect responded to the test mailing and made a purchase. It is likely a binary variable with values like 0 or 1, where 0 represents no purchase and 1 represents a purchase.
* Re\_update: This binary variable has been created from the reference of “1st\_update\_days\_ago” and “last\_update\_days\_ago” columns, where 1 means the update to the customer has been done more than once and 0 means update was only done once.
* Source\_unknown: the records that don’t belong to any sources or web. Orders are placed into this column.

**Structure of the data:**

|  |
| --- |
| Structure of the Data |
| 'data.frame': 2000 obs. of 26 variables: |
| $ US : int 1 1 1 1 1 1 1 1 1 1 ... |
| $ source\_a : int 0 0 0 0 0 0 0 0 1 1 ... |
| $ source\_c : int 0 0 0 1 1 0 0 0 0 0 ... |
| $ source\_b : int 1 0 0 0 0 0 0 1 0 0 ... |
| $ source\_d : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ source\_e : int 0 1 0 0 0 0 0 0 0 0 ... |
| $ source\_m : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ source\_o : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ source\_h : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ source\_r : int 0 0 0 0 0 1 0 0 0 0 ... |
| $ source\_s : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ source\_t : int 0 0 1 0 0 0 0 0 0 0 ... |
| $ source\_u : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ source\_p : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ source\_x : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ source\_w : int 0 0 0 0 0 0 1 0 0 0 ... |
| $ Freq : int 2 0 2 1 1 1 2 1 4 1 ... |
| $ last\_update\_days\_ago: int 3662 2900 3883 829 869 1995 1498 3397 525 … |
| $ X1st\_update\_days\_ago: int 3662 2900 3914 829 869 2002 1529 3397 2914 … |
| $ Web.order : int 1 1 0 0 0 0 0 0 1 0 ... |
| $ Gender.male : int 0 1 0 1 0 0 0 1 1 0 ... |
| $ Address\_is\_res : int 1 0 0 0 0 1 1 0 0 0 ... |
| $ Purchase : int 1 0 1 0 0 0 0 0 1 1 ... |
| $ Spending : int 128 0 127 0 0 0 0 0 489 174 ... |
| $ source\_unknown : int 0 0 0 0 0 0 0 0 0 0 ... |
| $ re\_update : int 0 0 1 0 0 1 1 0 1 0 ... |

**Attribute Analysis:**

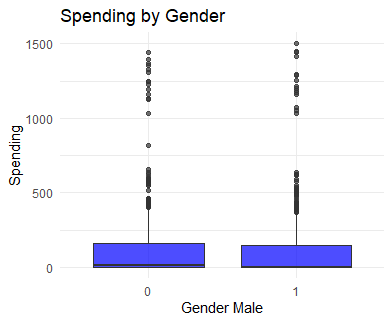
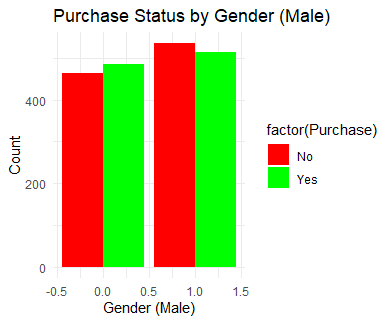
Scatter plot between Frequency and Spending variables.

****

**Fig 1**

The scatter plot above displays the relationship between Frequency and Spending, the majority of the data points are clustered at the lower end of the frequency scale which suggests most customers have low frequency on purchases. However, there is no clear upward or downward trend as the frequency increases. This suggests there's no linear relationship between the frequency of purchases and the amount spent.

Purchase and Spending status by Gender.

****

**Fig 2 Fig 3**

The above bar graph represents purchase status by gender, the count of males and females who made a purchase is roughly similar with a slightly higher count for males. The gender male has higher counts on both purchases and non-purchases, however, the difference between male and female purchases isn’t large.

The graph on the right represents spending by gender, both genders have similar median spending. The range of spending is broad for both genders.

Box plot between spending and Re-update variables.

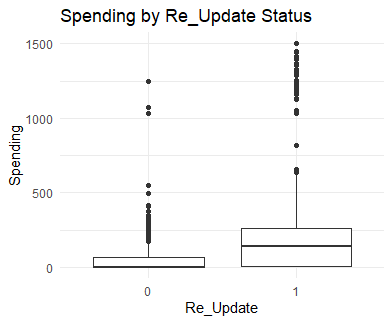


Fig 4

This boxplot shows the distribution of spending by the Re\_update status, which is a binary variable, where 1 means the update to the customer was done more than one time and 0 means the update to the customer was only done once.

The graph shows that the median spending for both groups is low on spending, however, the range of spending is high we can say that by looking at the outliers, especially for the customers who were updated more than once.

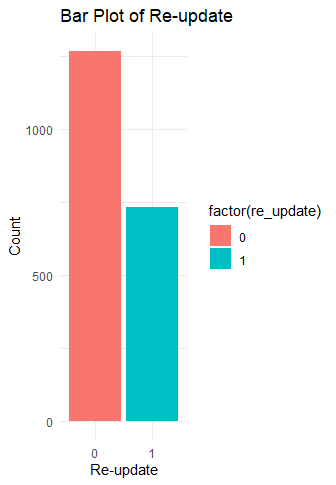
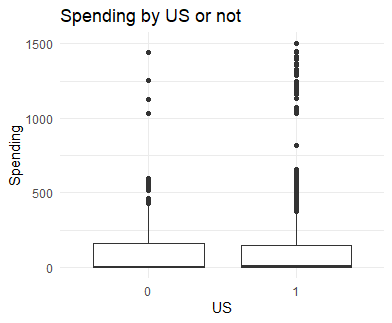


Fig 5

The bar graph shows the purchase count is higher for the customers who were updated only once.

Box plot between spending and marketplace US or not.

****

**Fig 6**

This box plot shows the distribution of spending by customer marketplace as US or not. graph represents median spending for both groups is similar, however, the outliers from the plot suggest that there are more instances of unusually high spending performed with US marketplace customers.

Bar plot between spending and all the sources

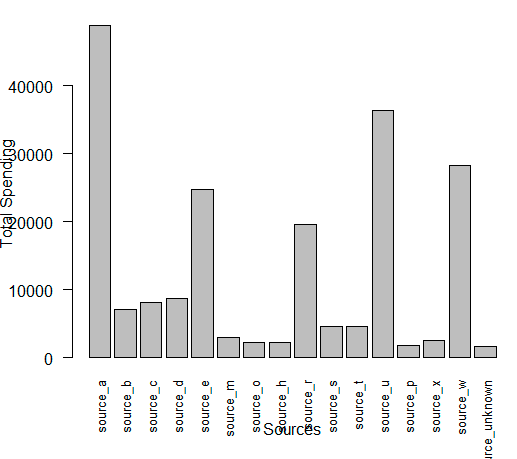
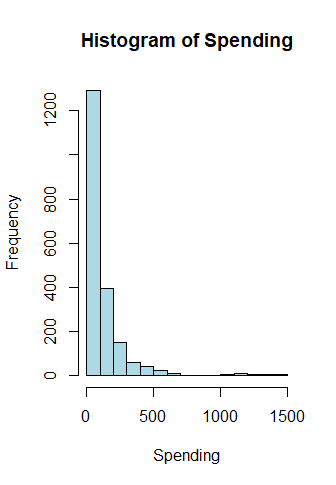


Fig 7

This graph represents the total sum of spending from each source in the data set from source\_a to source\_w.

Source\_a has the highest spending among all the sources, the spending from source\_e, source\_r, source\_u and source\_w is next highest.

Histogram for attribute spending

****

**Fig 8**

This graph shows that most spending amounts are low and large number of observations fall into lowest the spending interval (close to 0) and very few high spending amounts can be seen, indicating a right-skewed distribution.

Bar graph b/w address\_is\_res and Spending

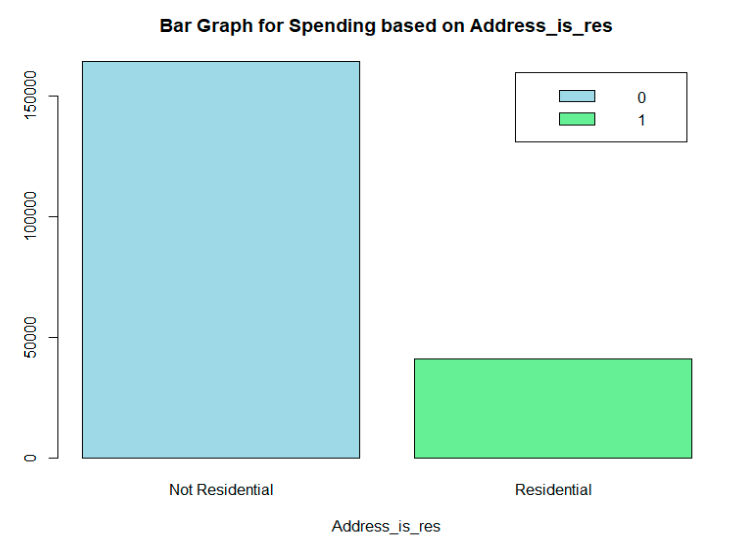


Fig 9

The bar graph shows that spending is higher for Non-residential addresses when compared to residential addresses.

Bar graph b/w address\_is res and Purchase

The bar graph represents that purchase count is higher for Non-residential address when compared to residential address

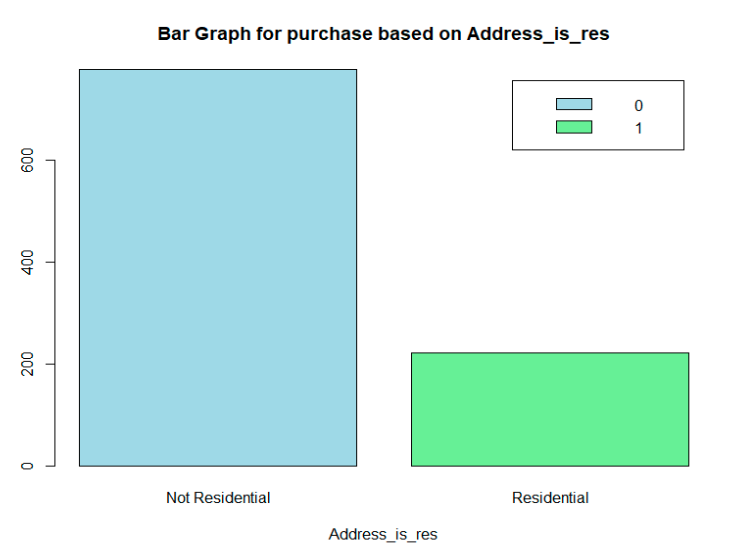


Fig 10

**Predictor analysis and Relevancy:**

This project involves the development of two machine learning models. The first model aims to classify whether a customer is a purchaser or not, while the second model focuses on predicting the amount of spending that a purchaser is likely to make.

Chi-square test Purpose:

The Chi-Square test is a statistical test used to determine if there is a significant association between two categorical variables, it is often employed to assess the relationship between predictor variables (categorical) and a categorical target variable.

Anova Purpose:

ANOVA is a statistical method used to assess whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. ANOVA is applied to numerical target variables and helps determine which predictor variables significantly contribute to the variance in the target.

Target Variable: Purchase

Chi-Square Test for Predictor Correlation:

The Chi-Square test was performed to assess the correlation between the predictor variables and the binary target variable 'Purchase'. The test results are summarized below:

|  |  |  |  |
| --- | --- | --- | --- |
| Target variable : Purchase | | | |
| Predictor variables | X-squared | df | p-value |
| Source\_a | 88.69 | 1 | < 2.2e-16 |
| Source\_c | 5.0034 | 1 | 0.0253 |
| Source\_b | 19.583 | 1 | 9.631E-06 |
| Source\_d | 1.257 | 1 | 0.2622 |
| Source\_e | 4.2473 | 1 | 0.03931 |
| Source\_m | 0.49298 | 1 | 0.4826 |
| Source\_o | 24.708 | 1 | 6.67E-07 |
| Source\_h | 64.33 | 1 | 1.052E-15 |
| Source\_r | 0.7836 | 1 | 0.376 |
| Souce\_s | 9.388 | 1 | 0.002184 |
| Source\_t | 4.6583 | 1 | 0.0309 |
| Source\_u | 46.743 | 1 | 8.093E-12 |
| Source\_p | 6.7907 | 1 | 0.009163 |
| Source\_x | 0.028287 | 1 | 0.8664 |
| Source\_w | 4.3173 | 1 | 0.03773 |
| Web.order | 112.92 | 1 | < 2.2e-16 |
| Gender.male | 0.80193 | 1 | 0.3705 |
| Address\_is\_res | 0.026139 | 1 | 0.8716 |
| re\_update | 301.23 | 1 | < 2.2e-16 |

Interpretation:

* A significant p-value (p < 0.05) suggests a strong association between the predictor and the target variable.
* Source\_a, Source\_h, Web.order, Source\_u, and re\_update show strong correlations with the likelihood of making a purchase.

Target Variable: Spending

ANOVA for Predictor Relevancy**:**

An ANOVA test was conducted to evaluate the relevance of the predictor variables for the numerical target variable 'Spending'. The results are summarized below:

|  |
| --- |
| Df Sum Sq Mean Sq F value Pr(>F) |
| US 1 845 845 0.057 0.811976 |
| Freq 1 33337504 33337504 2234.298 < 2e-16 \*\*\* |
| last\_update\_days\_ago 1 18454 18454 1.237 0.266229 |
| X1st\_update\_days\_ago 1 254826 254826 17.079 3.74e-05 \*\*\* |
| Web.order 1 179651 179651 12.040 0.000532 \*\*\* |
| Gender.male 1 14 14 0.001 0.975429 |
| Address\_is\_res 1 2163184 2163184 144.978 < 2e-16 \*\*\* |
| Purchase 1 3789394 3789394 253.967 < 2e-16 \*\*\* |
| re\_update 1 304348 304348 20.398 6.66e-06 \*\*\* |
| Residuals 1990 29692380 14921 |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

|  |
| --- |
| Df Sum Sq Mean Sq F value Pr(>F) |
| source\_a 1 2367119 2367119 72.784 < 2e-16 \*\*\* |
| source\_c 1 34284 34284 1.054 0.304678 |
| source\_b 1 127760 127760 3.928 0.047617 \* |
| source\_d 1 13419 13419 0.413 0.520723 |
| source\_e 1 41242 41242 1.268 0.260258 |
| source\_m 1 879 879 0.027 0.869397 |
| source\_o 1 282309 282309 8.680 0.003254 \*\* |
| source\_h 1 722087 722087 22.203 2.62e-06 \*\*\* |
| source\_r 1 195670 195670 6.016 0.014259 \* |
| source\_s 1 302325 302325 9.296 0.002327 \*\* |
| source\_t 1 39 39 0.001 0.972461 |
| source\_u 1 705532 705532 21.694 3.41e-06 \*\*\* |
| source\_p 1 54736 54736 1.683 0.194676 |
| source\_x 1 5194 5194 0.160 0.689484 |
| source\_w 1 395108 395108 12.149 0.000502 \*\*\* |
| source\_unknown 1 474 474 0.015 0.903891 |
| Residuals 1983 64492422 32523 |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

Interpretation:

* A significant p-value (p < 0.05) suggests that the predictor variable is relevant for explaining the variance in the target variable.
* Freq, Address\_is\_res, Purchase, re\_update, source\_a,source\_h,source\_u,source\_w are statistically significant predictors for explaining the variance in 'Spending'.

**Dimension Reduction:**

* The sequence variable can be excluded as it doesn’t contribute towards our analysis, and it just represents the row number.
* Excluding the new column created update\_status, which was introduced to check the relation between target variables, however it doesn’t have any significance with both of the target variables.

**Feature Selection using Stepwise Regression with AIC:**

Feature Selection for purchase model:

To enhance the predictive performance of our model and streamline the set of predictors, we employed a stepwise regression approach based on the Akaike Information Criterion (AIC). The goal was to identify a subset of features that maximally contributes to predicting the target variable, 'Purchase.' The model adds and removes predictors, evaluating the impact on the AIC at each step. The final model, with the lowest AIC value, suggests a better set of predictors for better performance. The selected features in our stepwise regression model for predicting 'Purchase' are shown below in table with the minimized AIC value of 1585.5.

|  |
| --- |
| Step: AIC=1585.47 |
| Purchase ~ source\_a + source\_d + source\_e + source\_m + source\_o + |
| source\_h + source\_r + source\_s + source\_t + source\_u + source\_p + |
| source\_x + source\_w + Freq + Web.order + Address\_is\_res |

|  |
| --- |
| Coefficients: |
| Estimate Std. Error z value Pr(>|z|) |
| (Intercept) -4.1828 0.2501 -16.725 < 2e-16 \*\*\* |
| source\_a 1.9567 0.2610 7.496 6.57e-14 \*\*\* |
| source\_d 0.5240 0.3483 1.505 0.132396 |
| source\_e 0.8574 0.2390 3.588 0.000333 \*\*\* |
| source\_m 1.5530 0.4983 3.117 0.001829 \*\* |
| source\_o 0.7174 0.4862 1.475 0.140099 |
| source\_h -3.4059 0.4576 -7.444 9.79e-14 \*\*\* |
| source\_r 1.2241 0.2920 4.192 2.76e-05 \*\*\* |
| source\_s 0.9739 0.3382 2.880 0.003979 \*\* |
| source\_t 1.2388 0.4238 2.923 0.003467 \*\* |
| source\_u 2.1136 0.2574 8.212 < 2e-16 \*\*\* |
| source\_p 2.4748 1.1278 2.194 0.028214 \* |
| source\_x 1.5422 0.4617 3.340 0.000837 \*\*\* |
| source\_w 1.3035 0.2324 5.609 2.03e-08 \*\*\* |
| Freq 2.4983 0.1379 18.121 < 2e-16 \*\*\* |
| Web.order 0.9208 0.1273 7.235 4.67e-13 \*\*\* |
| Address\_is\_res -0.6877 0.1764 -3.899 9.66e-05 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

Interpretation of Coefficients:

* source\_e, source\_m, source\_u: These sources positively contribute to the likelihood of a purchase, with statistically significant coefficients.
* source\_h: A one-unit increase in source\_h results in a significant decrease in the log-odds of a purchase (−3.4059), indicating a strong negative association.
* source\_r, source\_s, source\_t, source\_x, source\_w: Positive coefficients with statistical significance, indicating positive associations with purchase likelihood.
* Address\_is\_res: A negative coefficient (−0.6877) indicates a decrease in the log-odds of purchase when the address is residential.

Feature Selection for Spending model:

The selected features in our stepwise regression model for predicting 'Spending' are shown below in table with the minimized AIC value of 19194.36

|  |
| --- |
| Step: AIC=19194.36 |
| Spending ~ source\_a + source\_c + source\_d + source\_o + source\_h + |
| source\_r + source\_u + Freq + last\_update\_days\_ago + X1st\_update\_days\_ago + |
| Address\_is\_res + Purchase + source\_unknown + re\_update |

|  |
| --- |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) -21.470790 8.517816 -2.521 0.011790 \* |
| source\_a 16.203961 9.163942 1.768 0.077176 . |
| source\_c -36.665307 12.258618 -2.991 0.002815 \*\* |
| source\_d -22.738505 14.169015 -1.605 0.108696 |
| source\_o 38.309311 16.132598 2.375 0.017660 \* |
| source\_h -39.643792 14.747533 -2.688 0.007245 \*\* |
| source\_r 39.548288 11.103934 3.562 0.000377 \*\*\* |
| source\_u 18.669337 8.986377 2.078 0.037882 \* |
| Freq 80.588388 3.134489 25.710 < 2e-16 \*\*\* |
| last\_update\_days\_ago -0.020796 0.005763 -3.608 0.000316 \*\*\* |
| X1st\_update\_days\_ago 0.010754 0.005968 1.802 0.071733 . |
| Address\_is\_res -58.599997 7.494795 -7.819 8.59e-15 \*\*\* |
| Purchase 96.521997 6.795244 14.204 < 2e-16 \*\*\* |
| source\_unknown 29.253473 15.891054 1.841 0.065789 . |
| re\_update -29.751142 8.332000 -3.571 0.000364 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

Interpretation of Coefficients**:**

* source\_a, source\_o, source\_r, source\_u, Freq, last\_update\_days\_ago, Purchase, re\_update: These predictors have positive coefficients, indicating that an increase in these variables is associated with higher spending.
* source\_c, source\_d, source\_h, Address\_is\_res: Negative coefficients suggest a negative impact on spending. For instance, source\_c, source\_d, and source\_h indicate lower spending when they are present.
* X1st\_update\_days\_ago, source\_unknown: These predictors have positive coefficients but are not statistically significant at conventional significance levels.

**Data sampling:**

* random partitioning to create three key subsets: training, validation, and hold-out sets. The training set is used for building the model, the validation set is used to test the built model and the hold out set assess the performance of the built model.
* To ensure the reproducibility of results, a specific random seed (123) was set before performing the random partitioning, which allows to recreate the exact same data split in other models too.
* Below is the dimension for data sampling performed on the data set.
* Training Set: This set, consisting of 800 records.
* Validation Set: this set, consisting of 700 records.
* Hold-Out Set: This set, consisting of 500 records.

**Classification Models for Purchase:**

Logistic Regression Model:

The logistic regression model was built to predict whether a customer would make a purchase or not. choosing the predictors from stepwise regression performed with lowest AIC

* The model generates predicted probabilities for each observation in the validation dataset.
* Threshold: A threshold of 0.5 is set to convert predicted probabilities into binary predictions. If the predicted probability is above 0.5, the model predicts a purchase (1), otherwise, it predicts no purchase (0).
* Confusion Matrix: The confusion matrix summarizes the performance of the model by comparing predicted and actual values. In this case, the confusion matrix is as follows

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 266 | 79 |
| 1 | 67 | 288 |

It shows that there were 266 true negatives (correctly predicted no purchase), 288 true positives (correctly predicted a purchase), 79 false positives (predicted a purchase, but there was none), and 67 false negatives (failed to predict a purchase when there was one)

* The accuracy of the stepwise logistic regression model is approximately 79.14%, indicating that the model correctly predicted the outcome for about 79.14% of the observations in the validation dataset.

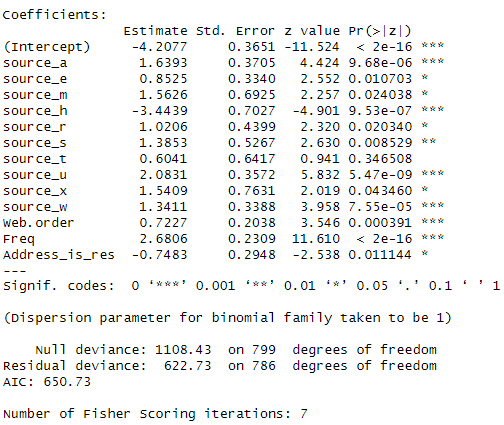


Fig 11

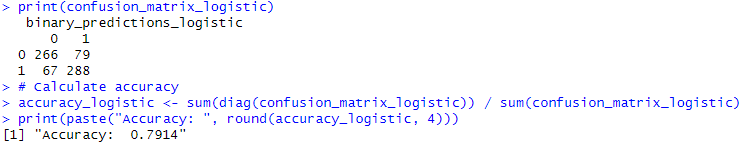


Fig 12

Classification Tree Model:

A classification tree is a predictive model that maps features (predictor variables) to a target variable (Purchase), typically used for classification tasks. The tree is constructed by recursively splitting the dataset based on the feature that provides the best separation according to a certain criterion (e.g., Gini index or entropy). Nodes represent decision points, and leaf nodes represent the predicted classes.

* Confusion Matrix: The confusion matrix summarizes the performance of the model by comparing predicted and actual values. In this case, the confusion matrix is as follows

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 271 | 74 |
| 1 | 71 | 284 |

It shows that there were 271 true negatives (correctly predicted no purchase), 284 true positives (correctly predicted a purchase), 74 false positives (predicted a purchase, but there was none), and 71 false negatives (failed to predict a purchase when there was one).

* Accuracy is the ratio of correctly predicted instances to the total instances. In this case, the model achieved an accuracy of approximately 79.29%, indicating that it correctly predicted the class for about 79.29% of the instances in the validation dataset

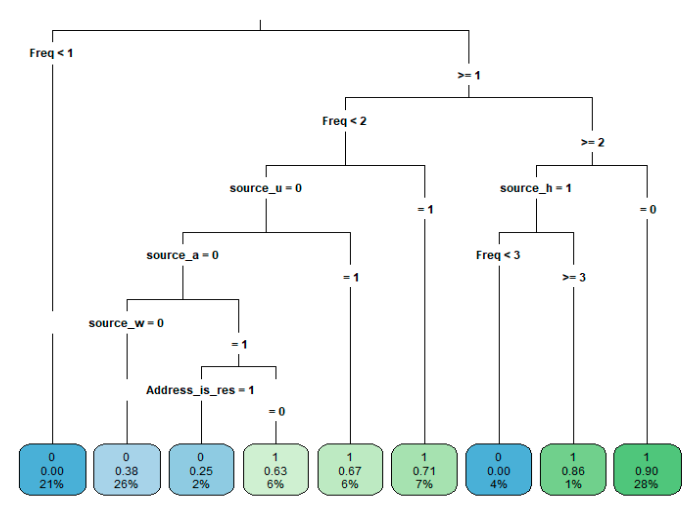


Fig 13

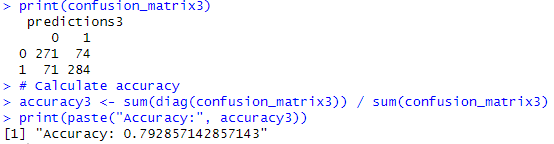


Fig 14

KNN model:

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for both classification and regression tasks. For our classification model, KNN makes predictions based on the majority class of the k-nearest data points in the feature space. The algorithm relies on the principle that similar instances tend to exist close to each other in the feature space.

* Confusion Matrix: The confusion matrix summarizes the performance of the model by comparing predicted and actual values. In this case, the confusion matrix is as follows

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 260 | 85 |
| 1 | 65 | 290 |

It shows that there were 260 true negatives (correctly predicted no purchase), 290 true positives (correctly predicted a purchase), 85 false positives (predicted a purchase, but there was none), and 65 false negatives (failed to predict a purchase when there was one).

* the KNN model achieved an accuracy of 78.57%, indicating that it correctly predicted the target variable (Purchase) for approximately 78.57% of the instances in the validation dataset.

Classification Model selection:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MODEL METRICS | | | | | |
| **Model** | **TN (Purchase = 0)** | **TP (Purchase=1)** | **FP (False Positive)** | **FN (False Negative)** | **Accuracy** |
| Logistic Regression | 266 | 288 | 79 | 67 | 79.14% |
| Classification Tree | 271 | 284 | 74 | 71 | 79.29% |
| K Nearest Neighbour | 260 | 290 | 85 | 65 | 78.57% |

Both the stepwise logistic regression and the classification tree models show similar accuracies, with the classification tree having a slightly higher accuracy. The KNN model lags in terms of accuracy compared to the other two models.

The business goal, presumably, is to accurately identify instances where a purchase is likely to occur. The logistic regression model demonstrates a higher true positive rate (288), indicating a better ability to correctly predict actual purchases. It also excels in this regard by minimizing false negatives (instances where a purchase is not predicted but occurs)

Therefore, the decision to choose logistic regression is not only based on statistical performance but also aligned with the specific business objective of identifying potential purchases.

**Regression Models for Spending:**

To enhance the focus on instances where a purchase has taken place and predict spending, subsets of the original datasets were created. These subsets exclusively contain records associated with purchasers only.

Multiple Linear Regression Model:

The multiple linear regression model was constructed to explore the relationship between spending behavior and several predictor variables. The model was built using training data focused on instances where purchases have occurred.

Coefficient Summary:

|  |
| --- |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 70.897614 26.269557 2.699 0.00727 \*\* |
| source\_r 74.552749 34.463985 2.163 0.03114 \* |
| Freq 98.088705 6.208863 15.798 < 2e-16 \*\*\* |
| Address\_is\_res -94.205492 22.805045 -4.131 4.44e-05 \*\*\* |
| last\_update\_days\_ago -0.014504 0.008645 -1.678 0.09422 . |
| re\_update -34.115796 21.859433 -1.561 0.11942 |

* source\_r Coefficient: The source\_r variable contributes significantly, with a coefficient of 74.55. This suggests that purchasers associated with source\_r tend to have higher spending.
* Address\_is\_res Coefficient: This variable has a negative coefficient of -94.21, suggesting that purchasers with a residential address tend to have lower spending.
* last\_update\_days\_ago Coefficient: The coefficient is -0.0145, indicating a marginal negative relationship between the time since the last update and spending.
* re\_update Coefficient: This variable has a negative coefficient of -34.12, suggesting a negative impact on spending for updated records.

Model Performance:



The model explains approximately 48.11% of the variability in spending, indicating a moderate level of predictive power. The adjusted R-squared, accounting for the number of predictors, is 47.43%.

Model Evaluation:



Root Mean Squared Error (RMSE): The RMSE for the model on the validation data is 148.34, representing the average difference between predicted and actual spending values.

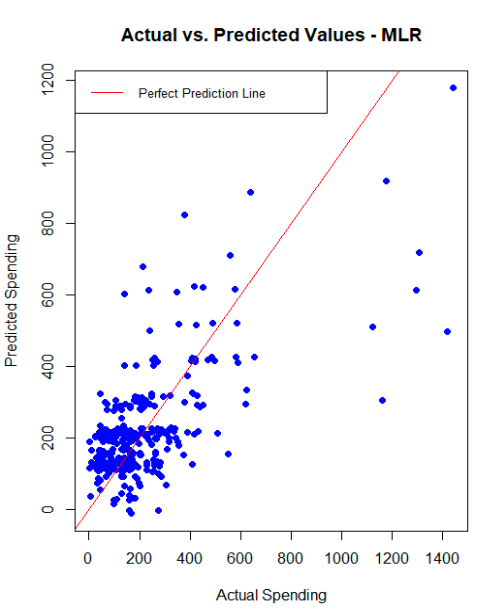


Fig 15

The scatter plot indicates a substantial linear relationship between the predicted and actual spending values.



The correlation between the actual and predicted spending values is 0.67. This strong positive correlation indicates that the model captures a considerable portion of the variability in spending.

Regression Tree Model:

regression tree used for predictive modeling, specifically in regression tasks. It is a decision tree-based algorithm that recursively partitions the data into subsets based on the values of input features, providing insights into the relationships between the predictors and the target variable (spending)

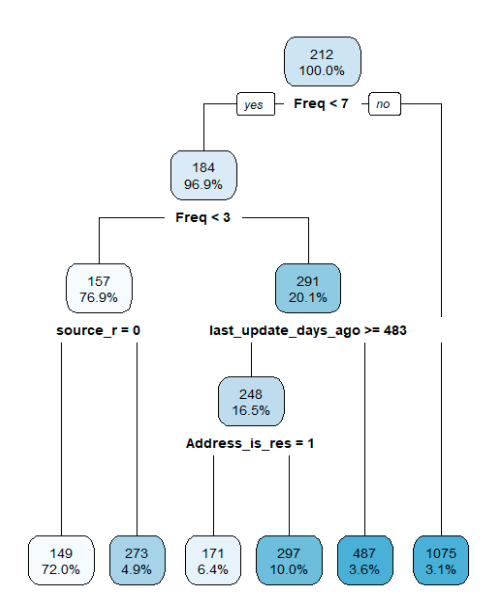
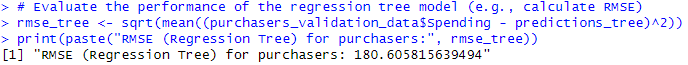


Fig 16

The root node (Node 1) has all 389 observations, employing primary splits based on features such as frequency (Freq), last\_update\_days\_ago, re\_update, source\_r, and Address\_is\_res. The mean spending at this node is 211.87, with a mean squared error (MSE) of 60241.85. Node 2 (Frequency < 6.5) comprises 377 observations, further partitioned by features, revealing a mean spending of 184.40 and an MSE of 32786.34. Node 3 (Frequency >= 6.5) . the observation goes through further more splits on different features .

Model Evaluation:

****

Root Mean Squared Error (RMSE): The RMSE for the regression tree model on the validation data is 180.61. This metric quantifies the average difference between predicted and actual spending values**.**

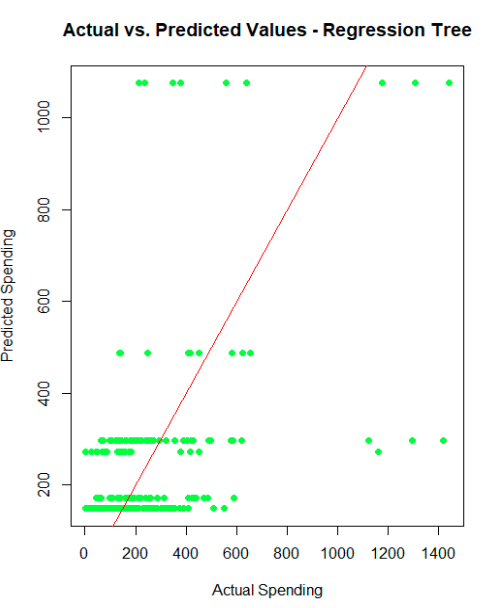
****

Fig 17

****

The correlation coefficient between the actual and predicted values obtained from the regression tree model is approximately 0.505. This indicates a moderate positive correlation between the predicted spending values generated by the regression tree and the actual spending values observed in the dataset.

Comparison between multiple linear regression and Regression tree model :

|  |  |  |
| --- | --- | --- |
| Metrics | | |
| Model | RMSE | Actual vs Predicted Correlation |
| Multiple Linear Regression | 148.34 | 0.67 |
| Regression Tree | 180.61 | 0.5 |

In evaluating the performance of the two predictive models for spending behavior, The MLR model yielded a Root Mean Squared Error (RMSE) of approximately 148.34 for purchasers, indicating the average deviation between predicted and actual spending values. Additionally, the correlation between the actual and predicted values was found to be approximately 0.667, signifying a moderately strong positive linear relationship.

Considering our business goal, which is to predict purchase likelihood and spending amounts, the MLR model appears to outperform the Regression Tree model. The MLR model demonstrates a lower RMSE and a stronger correlation.

**Model Evaluation on Holdout data**

Adding predicted\_probability column:

predictions were generated for the holdout dataset using the stepwise logistic regression model, which had been previously trained and validated. The logistic model aimed to predict the probability of a purchase for each observation in the holdout data. These predicted probabilities were then integrated into the holdout dataset as a new column labeled "Predicted\_Probability."

Adding Predicted\_spending column:

the multiple linear regression model, which was trained and validated using the available data, was employed to make predictions on the holdout dataset. The model was configured to estimate the spending values for each observation in the holdout data. Subsequently, the predicted spending values were integrated into the holdout dataset as a new column labeled "Predicted\_Spending."

Adding Adjusted\_probability\_purchase column:

an adjustment to the predicted purchase probabilities was performed based on a predefined original purchase rate assumption. The original purchase rate was set at 10.65% (reflecting 1065 purchases out of 20,000 samples with a 50% purchase rate). Subsequently, the predicted purchase probabilities, derived from the classification model, were scaled by this original purchase rate. The resulting values were then incorporated into the holdout dataset under the column labeled "Adjusted\_Probability\_Purchase."

Adding Expected\_spending Column:

the calculation of expected spending for the holdout data was conducted. The process involved multiplying the predicted spending values, generated from the multiple linear regression model, by the adjusted probability of purchase. The adjusted probability reflects the scaled probabilities considering a predefined original purchase rate assumption. The resulting values were then added to the holdout dataset under the column labeled "Expected\_Spending."

|  |
| --- |
| US source\_a source\_c source\_b source\_d source\_e source\_m source\_o source\_h source\_r source\_s source\_t |
| 1789 0 0 0 0 0 0 0 0 0 0 0 1 |
| 512 1 0 0 0 0 0 0 1 0 0 0 0 |
| 1433 1 1 0 0 0 0 0 0 0 0 0 0 |
| 1659 0 0 0 0 0 0 0 0 1 0 0 0 |
| 122 1 0 0 0 0 0 0 0 0 0 0 0 |
| 1752 1 0 1 0 0 0 0 0 0 0 0 0 |
| source\_u source\_p source\_x source\_w Freq last\_update\_days\_ago X1st\_update\_days\_ago Web.order |
| 1789 0 0 0 0 1 2880 2880 0 |
| 512 0 0 0 0 0 3747 3747 0 |
| 1433 0 0 0 0 5 1240 3620 0 |
| 1659 0 0 0 0 2 1275 1311 0 |
| 122 0 0 0 0 2 3436 3517 1 |
| 1752 0 0 0 0 1 1261 1261 1 |
| Gender.male Address\_is\_res Purchase Spending source\_unknown re\_update Predicted\_Probability |
| 1789 1 0 0 0 0 0 0.28433448 |
| 512 1 0 0 0 0 0 0.01466211 |
| 1433 1 0 1 271 0 1 0.99998029 |
| 1659 0 1 0 0 0 1 0.04570370 |
| 122 1 0 1 225 0 1 0.86716273 |
| 1752 1 0 1 44 0 0 0.30907445 |
| Predicted\_Spending Adjusted\_Probability\_Purchase Expected\_Spending |
| 1789 127.21405 0.030281622 3.8522477 |
| 512 16.55015 0.001561515 0.0258433 |
| 1433 509.24006 0.106497901 54.2329975 |
| 1659 120.26080 0.004867444 0.5853627 |
| 122 183.12259 0.092352831 16.9118893 |
| 1752 150.69645 0.032916429 4.9603888 |

**Gain Chart:**

Gain Chart was generated to evaluate the model's performance on the holdout data in predicting expected spending. The dataset was first sorted in descending order based on the predicted probability of spending, allowing for a systematic examination of cumulative expected spending.

The cumulative gain chart illustrates the cumulative sum of expected spending as more records are targeted. The x-axis represents the number of records targeted, while the y-axis showcases the cumulative expected spending. The blue reference line on the chart indicates the expected cumulative spending under random targeting, serving as a benchmark for comparison

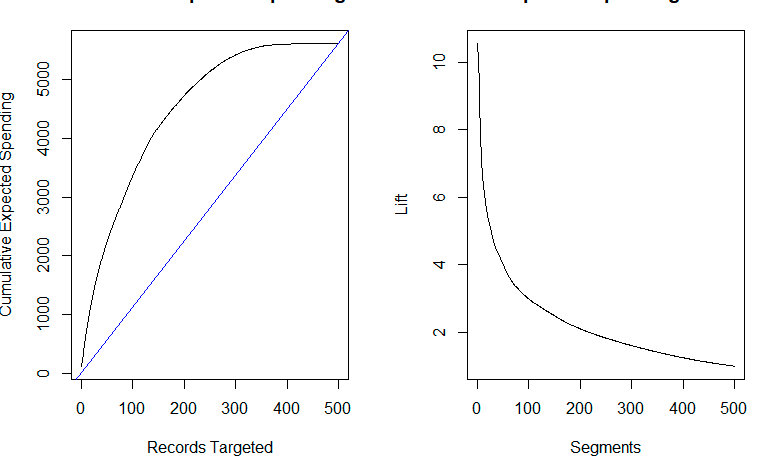


Fig 18

**Lift chart:**

Lift chart on the right measures how much better one can expect to do with predictive model compared to one without it. As we can see at the beginning of the curve the lift is highest , indicating that targeting the first segment of records identified by the model is much better than random targeting.

**Gross profit estimation:**

Calculating Estimated Response Rate for Remaining Customers:

Initial Response Rate: 5.3% (0.053)

Number of Remaining Customers: 180,000

Estimated Positive Responders = 180,000 \* 0.053 = 9,540

Calculating Gross profit per customer:



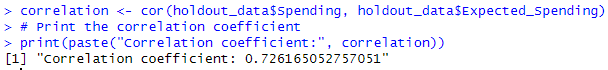
Gross profit per customer =Average expected spending - $2 for the mailing cost

=102.625

Total Gross Profit:

The overall gross profit for North-Point is determined by multiplying the estimated positive responders (9540) with the gross profit per customer, which is $102.625, resulting in $979,042.5. Consequently, this represents the anticipated outcome for the remaining 180,000 customers for the company North-Point.

Correlation between actual spending and expected spending :

****

The correlation coefficient between the actual spending and the predicted expected spending on the holdout data is calculated to be approximately 0.73. This strong positive correlation indicates a robust relationship between the model's predictions and the actual spending behavior.

**Business Recommendation:**

Model Selection:

For predicting the likelihood of purchases (classification), the Logistic Regression model is recommended. It demonstrates a higher true positive rate and minimizes false negatives, aligning with the business goal of accurately identifying potential purchases.

For predicting spending amounts (regression), the Multiple Linear Regression model is recommended. It exhibits a lower Root Mean Squared Error (RMSE) and a stronger correlation with actual spending values compared to the Regression Tree model.

Targeted Marketing Strategy:

Utilize the Logistic Regression model to identify potential purchasers with higher accuracy. This will optimize the selection of customer names from a large pool, maximizing the response rate in the mailing campaign for the new software listing.

Resource Optimization:

Leverage the Multiple Linear Regression model to predict spending amounts. This will help the company estimate the expected spending for each customer, allowing for more informed decisions on resource allocation and budgeting for the mailing campaign.

Source Analysis:

Investigate the sources contributing to the 'source\_unknown' category. Understanding and categorizing these sources can provide insights into potential new customer segments and improve the overall understanding of the customer base.

Update Frequency Analysis:

Further explore the 'Re\_update' variable and its impact on customer behavior. Understanding how and when customers are updated can provide valuable insights for targeted marketing efforts and customer engagement strategies

**Executive Summary**

North-Point, a software retail firm known for its games and educational software, has embarked on an initiative to expand its customer base by leveraging a consortium of listing firms. This initiative enables North-Point to access a shared pool of over 5 million customer names, enhancing its direct mailing campaigns for a newly revised product listing. The key business challenge is to optimize the selection process from this pool to maximize the response rate and the financial return on the mailing campaign.

To address this challenge, North-Point conducted a mailing experiment by selecting 20,000 names from the pool, achieving a response rate of 5.3%. The company then utilized a dataset comprising an equal number of purchasers and non-purchasers to develop two predictive models using machine learning techniques: one for predicting the likelihood of purchase and another for predicting the potential spending amount per customer.

The dataset included both categorical and numerical variables. Initial data exploration did not reveal any missing values but highlighted some discrepancies such as non-web order records and instances of spending without purchases. Advanced analytics techniques, including Chi-square and ANOVA tests, were applied to determine significant predictors for both models.

The logistic regression model identified key predictors for purchase likelihood with an accuracy of approximately 79.14%, while the multiple linear regression model provided insights into spending behaviors with significant variables like frequency of purchase and customer's geographic location (U.S. vs. non-U.S.).

The analysis suggests that North-Point can significantly enhance the effectiveness of its mailing campaign by using the developed models to target customers more precisely. By focusing on those most likely to purchase and those predicted to spend more, North-Point can increase its expected gross profit while reducing mailing costs.

This project demonstrates the power of data-driven decision-making in direct marketing campaigns. By effectively utilizing predictive modeling, North-Point is well-positioned to not only improve its marketing outcomes but also achieve a better understanding of customer behavior, which will aid in future strategic decisions.

USED SMARTPHONE PRICE PREDICTION & BUSINESS RECOMMENDATIONS

By Komal Singh

George Herbert Walker School of Business and Technology, Webster University

CSDA 6010: Data Analytics Practicum

Professor: Ali Ovlia

**Table of contents**

**Introduction……………………………………………………………………………………………**

**Business Goal ……………………………………………………………………………**

**Dataset …………………………………………………………………………………………**

**Variable Description ………………………………………………………………………………….**

**Missing values ………………………………………………………………………………………**

Handling missing values …………………………………………………………………………..

**Data Transformation …………………………………………………………………………………….**

**Binary Conversion ………………………………………………………………………………**

**New Variable Introduction ………………………………………………………………………………**

Device Type ………………………………………………………………………………………

Thickness …………………………………………………………………………………………

Price Drop ………………………………………………………………………………………

**Data Structure after Preprocessing ………………………………………………………………**

**Feature Analysis ………………………………………………………………………………………**

Scatter plot for days\_used vs Used price …………………………………………………………

Scatter plot front and rear cameras vs used price: …………………………………………

Scatter plot screen size vs used price: ……………………………………………………………

Scatter plot battery vs used price: ………………………………………………………………

Scatter plot for release year vs used price: ………………………………………………………

Scatter plot of used price vs different phone brands : …………………………………………

**Finding Competitors ……………………………………………………………………………………**

**Analysis of sales revenue and usage patterns …………………………………………………**

**Analysis of phone models by brand and year ……………………………………………………**

**Operating system preferences ……………………………………………………………………………**

**Analysis of tablet distribution ………………………………………………………………………**

**Thickness analysis of tablets ………………………………………………………………………**

**Camera analysis ……………………………………………………………………………………**

**Nokia product analysis ………………………………………………………………………………**

Product lifecycle …………………………………………………………………………………

Nokia OS distribution over years …………………………………………………………………

**Feature selection and relevancy ……………………………………………………………………**

**Data partition ………………………………………………………………………………**

Training set …………………………………………………………………………………

Testing set ………………………………………………………………………………………

Holdout set …………………………………………………………………………………………

**Predictive modeling for used phone prices ……………………………………………………………**

Model development ………………………………………………………………………………

Model evaluation …………………………………………………………………………………

**Conclusion …………………………………………………………………………………………**

Recommendations for Nokia ………………………………………………………………………

Predictive model selection …………………………………………………………………………

**References ……………………………………………………………………………………………**

**Executive Summary**

**Introduction:**

This project investigates used mobile phone pricing with an emphasis on the wide range of variables impacting consumer decision. You will be analyzing how phone features such as screen size, rear and front camera, internal memory, and other properties affect the price of a used phone. You start with exploratory data analysis on the features to explore the relationship between the explanatory variables and the response (target) variable, You also used this information to identify possible multi-collinearity and identify potential outliers in the dataset. You explore the mathematical relationship between the explanatory variables and response variable.

Project

The modern mobile phone market is a complex ecosystem where pricing strategies are influenced by a multitude of factors. Predictive models such as regression are essential for predicting the prices of mobile phones, or classification models can classify the price such as high or low which, will help providers of new or used phone as well as customers in informed decision-making.

This project seeks to explore the complex correlations between fundamental phone features and their impact on new or used mobile phone prices. This study aims to uncover the crucial determinants shaping pricing strategies within the mobile phone industry. The first core task of the project requires data explanation and understanding. The second task of this project is on explorative data analysis where you explore some relationship, trends, pattern and correlation in the variables . The third task of the project is the data preparation for modeling. The fourth task you will build predictive and/or classification model using selected explanatory variables, and then model results analysis. In the final section of this project will be the conclusions about your findings and discuss further work on how to improve the model to be more robust and more accurate (recommendations)

**Business Goal:**

Our business goal is to leverage comprehensive data analytics on Nokia used devices, aiming to find pivotal factors influencing pricing and sales trends. By conducting in-depth feature comparisons with competitors, examining usage patterns, exploring the product life cycle, and scrutinizing operating system preferences, we strive to provide actionable insights for Nokia. Additionally, our predictive analysis on used prices seeks to empower Nokia with foresight into market dynamics. The overarching objective is to equip Nokia with strategic recommendations for enhancing sales performance, identifying areas of improvement, and ensuring a competitive edge in the dynamic mobile device market. Through these endeavors, we aim to contribute to Nokia's success by fostering informed decision-making and customer-centric product enhancements.

**Dataset:**

dataset indicates that it consists of 3454 observations (rows) and 15 variables (columns)

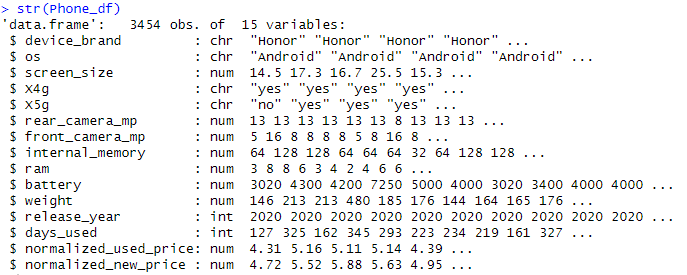
****

Fig: 1

**Variables description:**

*device\_brand***:** Character variable representing the brand of the mobile device (e.g., "Honor").

*os:* Character variable denoting the operating system of the mobile device (e.g., "Android").

*screen\_size:* Numeric variable indicating the size of the device's screen.

*X4g:* Character variable indicating the presence of 4G connectivity (e.g., "yes" or "no").

*X5g:* Character variable indicating the presence of 5G connectivity (e.g., "yes" or "no").

*rear\_camera\_mp:* Numeric variable representing the megapixels of the rear camera.

*front\_camera\_mp:* Numeric variable representing the megapixels of the front camera.

*internal\_memory:* Numeric variable indicating the internal storage capacity of the device.

*ram:* Numeric variable representing the Random Access Memory (RAM) capacity of the device.

*battery:* Numeric variable indicating the battery capacity of the device.

*weight:* Numeric variable representing the weight of the device.

*release\_year:* Integer variable indicating the year when the device was released.

*days\_used:* Integer variable representing the number of days the device has been used.

*normalized\_used\_price:* Numeric variable representing the normalized price of the used device.

*normalized\_new\_price:* Numeric variable representing the normalized price of a new device.

**Missing values:**

there are a total of 202 missing values across all observations. The missing values are distributed across several columns. Here's the breakdown of missing values for each column in below table.

|  |  |
| --- | --- |
| Column Name | Missing Values |
| device\_brand | 0 |
| os | 0 |
| screen\_size | 0 |
| X4g | 0 |
| X5g | 0 |
| rear\_camera\_mp | 179 |
| front\_camera\_mp | 2 |
| internal\_memory | 4 |
| ram | 4 |
| battery | 6 |
| weight | 7 |
| release\_year | 0 |
| days\_used | 0 |
| normalized\_used\_price | 0 |
| normalized\_new\_price | 0 |
| Total | 202 |

**Handling Missing Values:**

handling missing values by imputing them based on similar features within the same brand

Identifying similar phones:For each phone with missing values, identify other phones from the same brand that have complete information for the missing variable.

Calculate Average or Median: Calculate the average or median of the non-missing values for the similar phones. This will be used as the imputed value for the missing variable.

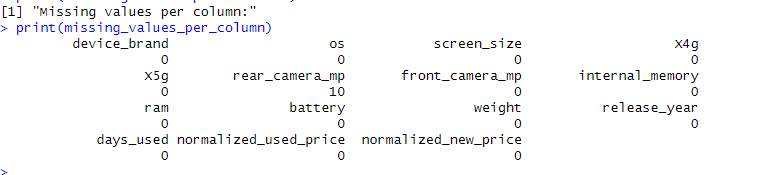
****

Fig: 2

After completing the imputation process, we still have 10 missing values in the 'rear\_camera\_mp' column as we can see from fig 2. This is attributed to the absence of comparable data for the brand "Infinix,"

The targeted imputation strategy initially aimed to calculate the average 'rear\_camera\_mp' separately for 'front\_camera\_mp' values of 8mp and 16mp within the specified years (2019 and 2020). While this approach successfully filled missing values (results in fig 3) for Infinix .

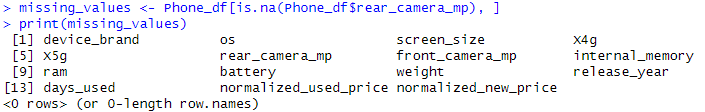


Fig: 3

**Data Transformation:**

|  |
| --- |
| Release year Android Others Windows |
| 2013 0 12 10 |
| 2014 5 4 11 |
| 2015 0 6 0 |
| 2016 0 2 0 |
| 2017 6 2 0 |
| 2018 12 2 0 |
| 2019 13 5 0 |
| 2020 10 6 0 |

In this section, we discuss the transformation applied to the dataset to ensure consistency and accuracy in our analysis. One notable transformation involved updating the operating system (OS) types for a subset of phones within the dataset.

Upon reviewing the data, it was observed that there were 15 phones of the brand Nokia released between 2017 and 2020 that were initially labeled with the OS type "Others." Our analysis indicated that Nokia exclusively adopted the Android operating system for its devices after the year 2017. Therefore, to maintain consistency and align with the known practices of Nokia, we made the decision to update the OS type of these 15 phones from "Others" to "Android."

This data transformation was performed with a specific focus on the Nokia brand. As our business goal is to provide actionable insights tailored to Nokia

**Binary Conversion:**

The columns 'X4g' and 'X5g,' which originally denoted the presence or absence of 4G and 5G capabilities, were converted into binary format. For both columns, the values 'yes' were replaced with '1,' indicating the presence of the feature, while 'no' was replaced with '0,' signifying the absence.

Additionally, we expanded the transformation to the 'os' column, representing the operating system of the mobile phones. Instead of a single column, we created separate binary columns for each unique value in the 'os' column. This was achieved using the model.matrix function.

**New variable introduction :**

Device type:

The bar chart in fig:4 illustrates that the predominant category in our dataset comprises phones, while tablets represent a minority. This classification is based on a threshold of 18 centimeters, marking the starting size for tablets.

we introduce a new variable, "device\_type," which was incorporated into the dataset to provide insights into the classification of products as either phones or tablets. The decision to introduce this variable stemmed from the need to categorize devices based on their form factor and screen size, allowing for a more granular analysis of the dataset.

The "device\_type" variable is derived from a reference point of 18 centimeters, which is commonly used as a threshold to distinguish between smartphones and tablets [1]. Devices with screen sizes below 18 centimeters are categorized as "Phone," while those with screen sizes equal to or greater than 18 centimeters are categorized as "Tablet." We have total of 3185 phones and 269 tablets in our dataset.

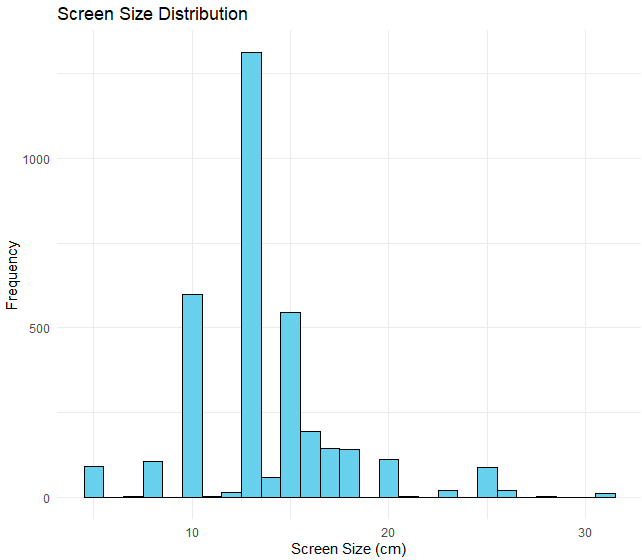


Fig: 4

Thickness:

we introduce a new variable, "thickness," which represents the estimated thickness of the device based on the screen size and weight. The decision to introduce this variable stemmed from the need to provide additional insights into the physical characteristics of tablets within our dataset.

Thickness was calculated specifically for tablets, as they typically have a uniform shape and lack physical components such as keypads that may affect the accuracy of the calculation. Tablets are characterized by their larger screen sizes, allowing for a more reliable estimation of thickness based on weight and screen area.

Conversely, the calculation of thickness was not extended to phones within our dataset. This decision was made due to the variability in phone designs, which can include smaller screen sizes and additional physical components such as keypads. As a result, accurately estimating the thickness of phones based solely on weight and screen size would introduce significant uncertainty and potential inaccuracies.

*Methodology****:***

Assuming the devices are in shape of a rectangle

Thickness = (Weight / Screen area)

Screen Area = Screen size \* Screen size

Price Drop:

The 'price\_drop' column has been added to our dataset, capturing the numerical representation of the decline in price for each mobile phone. The price drop is calculated by subtracting the normalized used price from the normalized new price. This metric provides valuable insights into the reduction in value that occurs as phones transition from being new to being used.

**Structure of data after preprocessing:**

|  |
| --- |
| 'data.frame': 3454 obs. of 21 variables: |
| $ device\_brand : chr "Honor" "Honor" "Honor" "Honor" ... |
| $ screen\_size : num 14.5 17.3 16.7 25.5 15.3 ... |
| $ X4g : num 1 1 1 1 1 1 1 1 1 1 ... |
| $ X5g : num 0 1 1 1 0 0 0 0 0 0 ... |
| $ rear\_camera\_mp : num 13 13 13 13 13 13 8 13 13 13 ... |
| $ front\_camera\_mp : num 5 16 8 8 8 8 5 8 16 8 ... |
| $ internal\_memory : num 64 128 128 64 64 64 32 64 128 128 ... |
| $ ram : num 3 8 8 6 3 4 2 4 6 6 ... |
| $ battery : num 3020 4300 4200 7250 5000 4000 3020 3400 4000 4000 ... |
| $ weight : num 146 213 213 480 185 176 144 164 165 176 ... |
| $ release\_year : int 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ... |
| $ days\_used : int 127 325 162 345 293 223 234 219 161 327 ... |
| $ normalized\_used\_price: num 4.31 5.16 5.11 5.14 4.39 ... |
| $ normalized\_new\_price : num 4.72 5.52 5.88 5.63 4.95 ... |
| $ device\_type : chr "Phone" "Phone" "Phone" "Tablet" ... |
| $ thickness : num NA NA NA 0.738 NA ... |
| $ price\_drop : num 0.408 0.357 0.774 0.496 0.558 ... |
| $ Android : num 1 1 1 1 1 1 1 1 1 1 ... |
| $ iOS : num 0 0 0 0 0 0 0 0 0 0 ... |
| $ Others : num 0 0 0 0 0 0 0 0 0 0 ... |
| $ Windows : num 0 0 0 0 0 0 0 0 0 0 ... |

After cleaning and pre processing the data we have a total of 21 variables in our dataset.

**Feature analysis:**

Scatter plot for days\_used vs Used price:

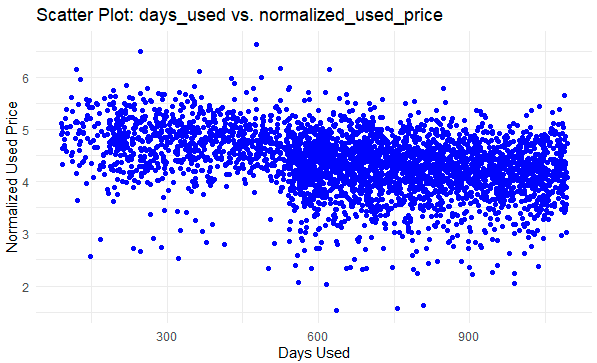


Fig:5

The provided scatter plot in fig 5 illustrates the distribution of used prices based on the number of days a phone has been in use. It reveals a gradual decline in prices as the duration of phone usage increases.

Scatter plot front and rear cameras vs used price:

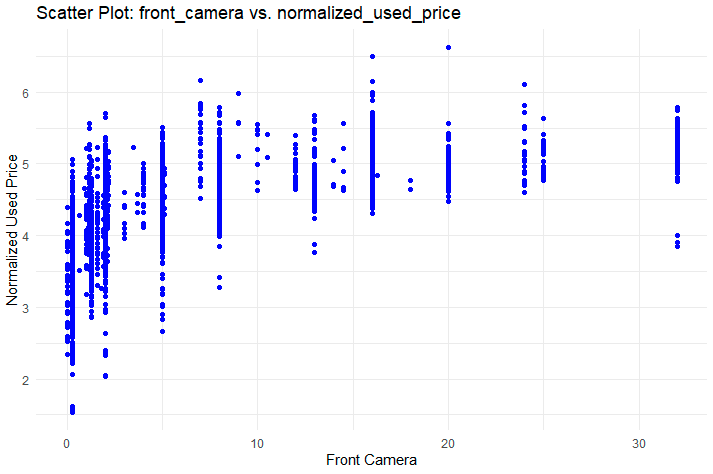


Fig:6

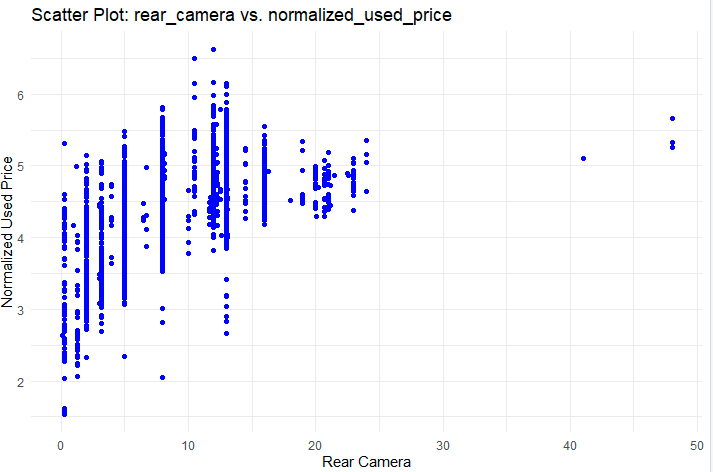


Fig: 7

The scatter plots in fig:6 and fig:7 depict the distribution of used prices concerning front and rear camera megapixels. The graphs highlight that phones with lower megapixels on their front and rear cameras tend to have lower prices compared to those equipped with higher megapixel cameras.

Scatter plot screen size vs used price:

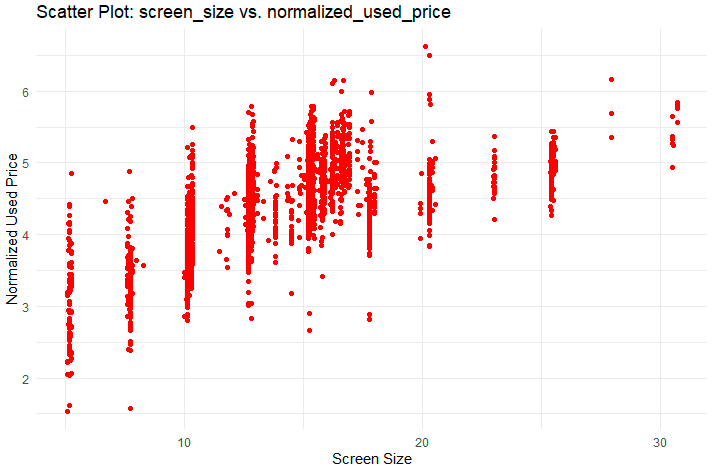


Fig: 8

The scatter plot in fig:8 showcases the distribution of used prices in relation to phone screen sizes. The plots indicate that phones with smaller screen sizes generally have lower prices.

Scatter plot battery vs used price:

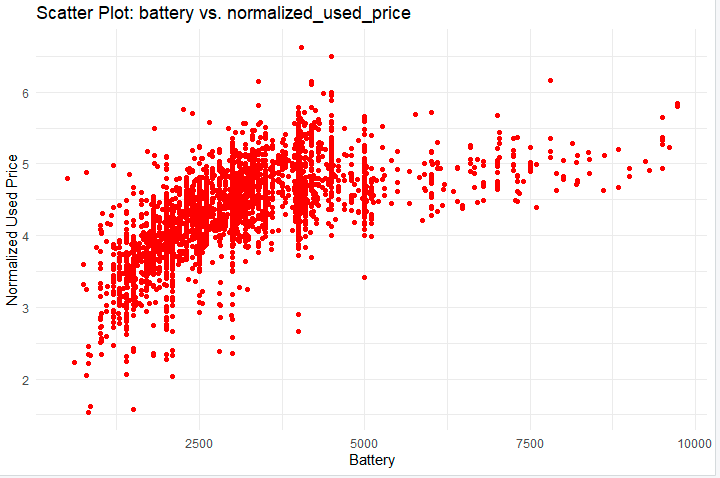


Fig: 9

The scatter plot in fig:9 displays the distribution of used prices based on phone battery specifications. Phones with batteries exceeding 5000 mAh are more likely to have prices above 4.

Scatter plot for release year vs used price:

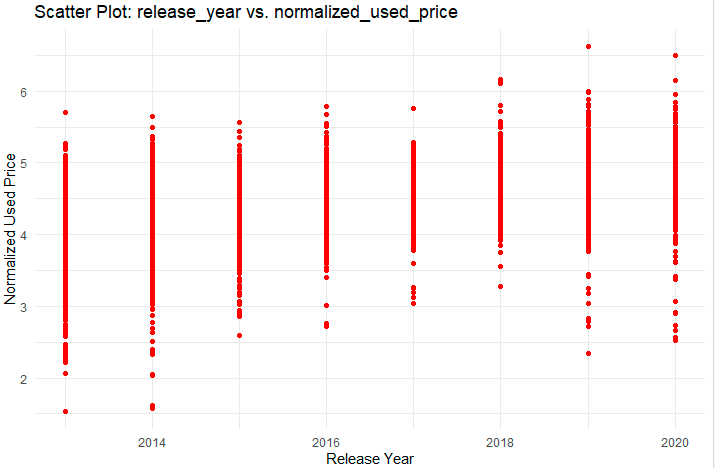


Fig:10

The scatter plot in fig:10 illustrates the distribution of used prices relative to the year of phone release. The graph reveals that phones released in 2017 exhibit less variation in prices, ranging only between 3-6. In contrast, phones released in 2013 and 2014 display a wider price variance from 1-6. The highest prices for used phones are observed in the years 2019 and 2020.

Scatter plot of used price vs different phone brands :

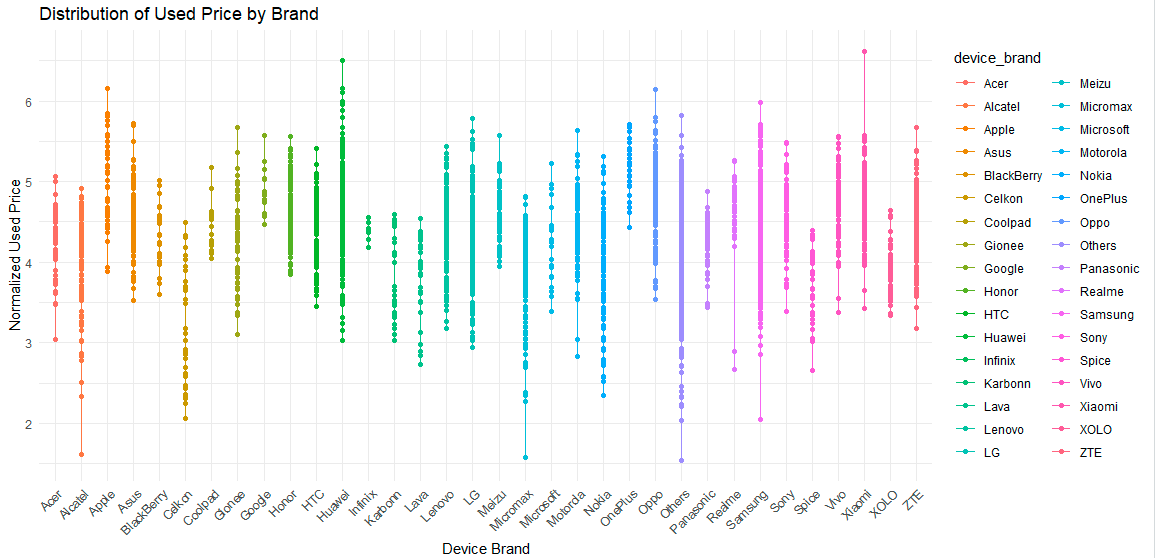


Fig:11

The scatter plot in fig:11 visualizes the distribution of used prices across different phone brands. It indicates that brands like Apple, Huawei, Oppo, and Xiaomi have the highest-priced phones, while brands like AI Catel, Celkon, Micromax, and others offer phones with lower prices.

**Finding Competitors:**

analysis aimed at identifying competitors to the Nokia brand within the market, gauged by the count of phones in our dataset. The top five competitors with the highest counts, in descending order, are Samsung, Huawei, LG, Lenovo, and ZTE.

The graph in fig:12 shows all the competitor brands for Nokia which had higher sales than nokia. Highlighted in blue

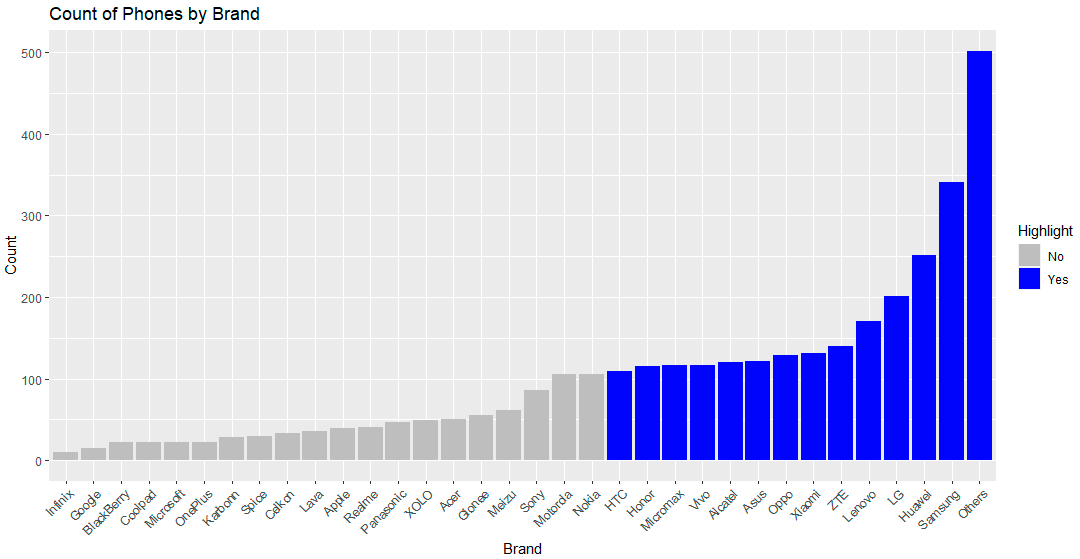
****

Fig: 12

**Analysis of Sales Revenue and Usage Patterns:**

The analysis focuses on comparing the sales revenue and usage patterns of Nokia with its key competitors in the mobile phone market. The dataset provides insights into sales revenue for both new and used phones, along with the corresponding revenue drop percentage. Additionally, it includes information on the maximum and minimum days phones are used for, offering valuable insights into consumer behavior and product lifecycle.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Device brand | Sales Revenue New | Sales Revenue used | Revenue drop % | Most days used | Least days used |
| Nokia | 494.8106 | 414.651 | 16.20 | 1093 | 127 |
| samsung | 1869.666 | 1523.107 | 18.53 | 1093 | 98 |
| Huawei | 1384.286 | 1168.296 | 15.60 | 1093 | 91 |
| LG | 1062.62 | 863.6885 | 18.72 | 1093 | 92 |
| Lenovo | 878.677 | 748.5333 | 14.76 | 1093 | 127 |
| ZTE | 730.6131 | 611.5333 | 16.29 | 1093 | 127 |

Sales Revenue:

New vs. Used Sales Revenue: Nokia generates $494.8106 in sales revenue from new phones and $414.651 from used phones, resulting in a revenue drop percentage of 16.20%.

Comparatively, Samsung, Huawei, LG, Lenovo, and ZTE exhibit varying revenue figures for both new and used phones, with revenue drop percentages ranging from 14.76% to 18.72%.

Usage Patterns:

Most Days Used: All brands, including Nokia, Samsung, Huawei, LG, Lenovo, and ZTE, have a consistent maximum usage duration of 1093 days.

Least Days Used: Nokia, Lenovo, and ZTE have the same minimum usage duration of 127 days, while Samsung, Huawei, and LG demonstrate relatively shorter minimum usage durations, ranging from 91 to 98 days.

Insights:

Price Drop Differences: While Nokia exhibits a moderate revenue drop percentage of 16.20%, LG has the highest revenue drop percentage at 18.72%, indicating a comparatively faster decline in value over time.

Least Days Used Comparison: Nokia, Lenovo, and ZTE share similar minimum usage durations of 127 days, suggesting relatively stable product usage among these brands. In contrast, Samsung, Huawei, and LG show shorter minimum usage durations, which may indicate a quicker turnover of devices or lower user satisfaction levels.

**Analysis of Phone Models by Brand and Year:**

In this section, we analyze the number of phone models released by various brands over the years. For each brand and year, we counted the distinct phone models based on specific features such as screen size, 4G/5G support, camera resolution.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Brand | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
| Samsung | 36 | 40 | 22 | 18 | 18 | 18 | 28 | 17 |
| Huawei | 19 | 19 | 14 | 13 | 13 | 19 | 21 | 10 |
| LG | 17 | 26 | 17 | 15 | 7 | 11 | 13 | 7 |
| Lenovo | 20 | 24 | 19 | 16 | 6 | 8 | 8 | 0 |
| ZTE | 15 | 14 | 22 | 20 | 13 | 10 | 12 | 3 |
| Nokia | 15 | 13 | 3 | 1 | 7 | 11 | 12 | 5 |

Samsung consistently leads in the number of phone models released each year, indicating a wide range of offerings to cater to diverse consumer preferences. Huawei also maintains a significant presence in the market, with a consistent number of phone models released each year. LG's phone model releases show some variability over the years, with peaks and dips in certain years. Lenovo and ZTE These brands exhibit a declining trend in the number of phone models released in recent years,

**Operating System Preferences:**

Examining the distribution of operating systems (Android, iOS, Windows, Others) for all devices and devices from brand Nokia.

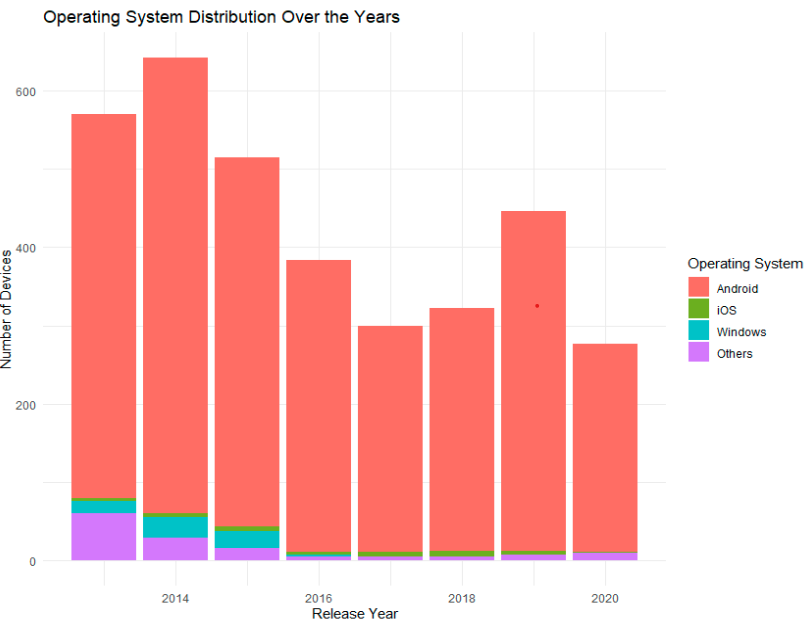


Fig:13

The bar chart in fig:13 illustrates the evolution of operating system usage across different phone brands over the years. Notably, post-2016, there is a prominent shift towards the Android OS, indicating a prevalent preference among brands. Conversely, the utilization of Windows OS appears to have declined in the later years, suggesting a decreasing trend in its adoption.

**Analysis of Tablet Distribution:**

The bar plot in fig:14 illustrates the distribution of tablets among Nokia and competitor brands. From the graph, it is evident that LG, Samsung, and ZTE exhibit a higher count of tablets compared to Nokia, Huawei, and Lenovo. This observation suggests a disparity in the emphasis placed on tablet production across different brands in the market.

While LG, Samsung, and ZTE seem to prioritize tablet offerings as part of their product strategy, Nokia, Huawei, and Lenovo appear to have a lesser focus on this particular device category.

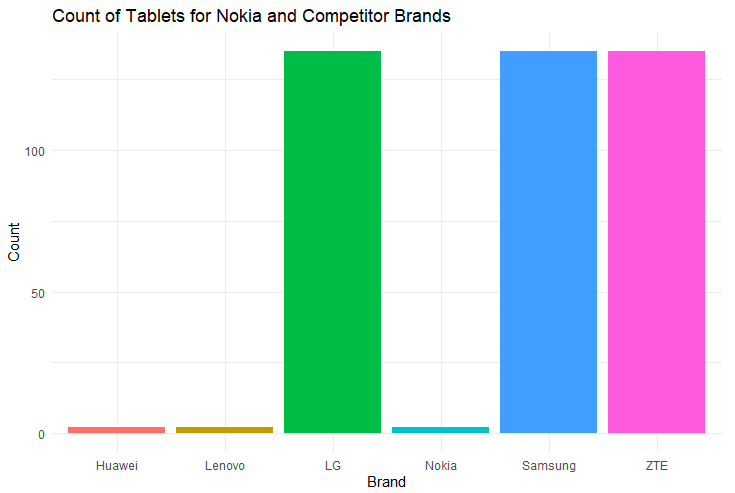
****

Fig: 14

**Thickness Analysis of tablets:**

In our analysis of tablet thickness for Nokia and competitor brands, we examined the minimum and maximum thickness values to gain insights into the design variations among different tablets

|  |  |  |
| --- | --- | --- |
|  | Thickness | |
| Brand | Minimum | Maximun |
| Nokia | 0.951 | 0.98 |
| Samsung | 0.201 | 1 |
| Huawei | 0.479 | 0.958 |
| LG | 0.346 | 0.84 |
| Lenovo | 0.724 | 1.131 |

These results highlight the variability in tablet thickness across different brands. Nokia tablets exhibit a relatively consistent thickness range, while competitor brands show a wider variation, ranging from 0.201 mm to 1.131 mm. Understanding these design differences can provide valuable insights for consumers and manufacturers alike, influencing purchasing decisions and product development strategies.

**Camera Analysis:**

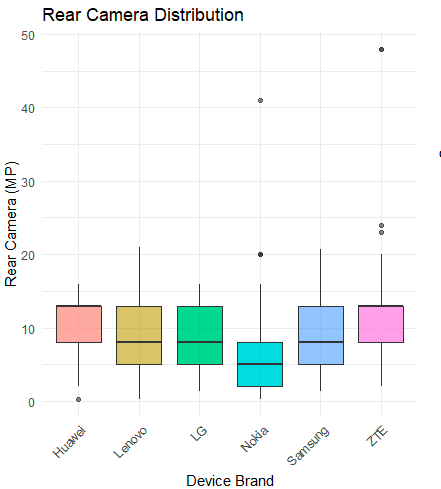
****

Fig:15

For Nokia, the box plot in fig:15 suggests that The median rear camera resolution is around 12-13 MP, which is in line with several other competitors. The interquartile range (IQR), representing the middle 50% of the data, is relatively tight, indicating that most of Nokia's devices have camera resolutions that do not vary widely. Compared to competitors, Nokia does not have the highest median resolution, but it also doesn't have the lowest. Its distribution is similar to that of Lenovo and LG, suggesting they offer comparable camera resolutions.

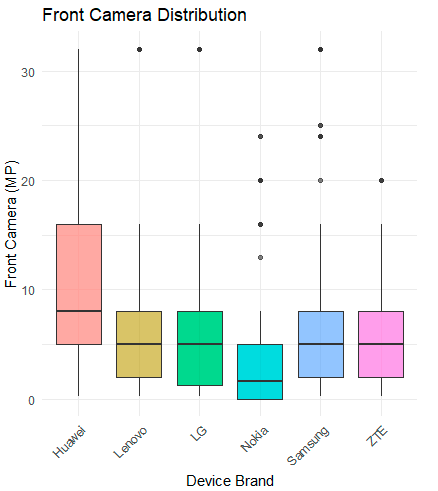


Fig:16

The box plot in fig:16 suggests that the median front camera resolution for Nokia devices is around 5 MP, which is lower than that of Huawei and Samsung but on par with Lenovo and LG.The interquartile range for Nokia is broader than for its rear cameras, indicating more variability in the front camera resolutions of Nokia devices.In comparison with other brands, Nokia's front camera offerings demonstrate a focus on a balance between quality and cost, potentially catering to a consumer base that values practicality over high-end specifications.

**Nokia Product analysis:**

Product Life Cycle:

Exploring the release year of Nokia devices and determining if there's a pattern indicating the introduction of new models and if older models are still relevant.

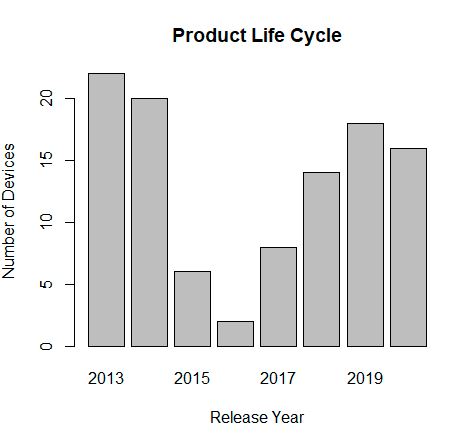


Fig:17

In the analysis of Nokia's product life cycle, from the graph in fig:17, it is evident that the years 2015, 2016, and 2017 saw a comparatively lower number of device releases. This observation suggests a potential dip or slowdown in Nokia's product output during this period. Factors influencing this trend could include market dynamics, technological transitions, or strategic shifts within the company.

Nokia Operating systems distribution over years:

The chart in fig:18 illustrates the evolution of operating systems used by Nokia over the years. Notably, up to 2016, Nokia predominantly employed 'Others' and 'Windows' as operating systems. However, post-2016, a significant shift occurred with Nokia adopting Android, leading to a gradual increase in the number of devices utilizing this operating system.

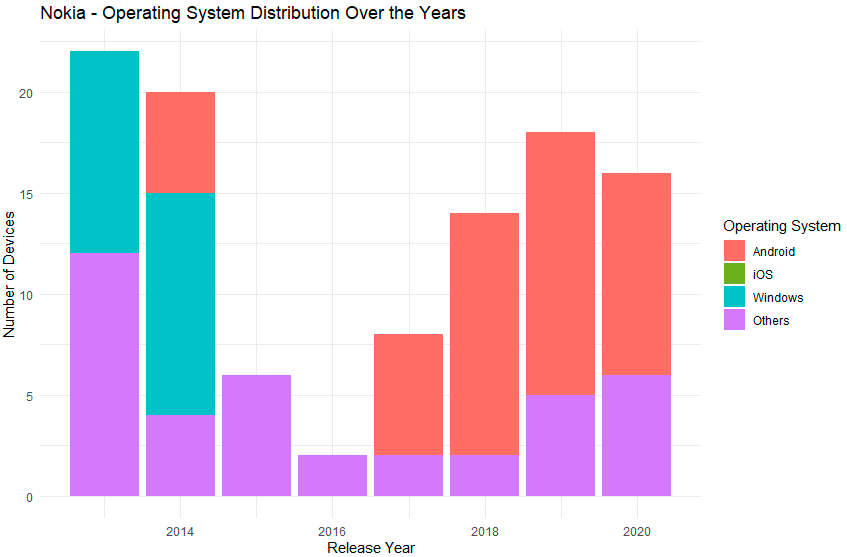


Fig:18

**Feature Selection and Relevancy:**

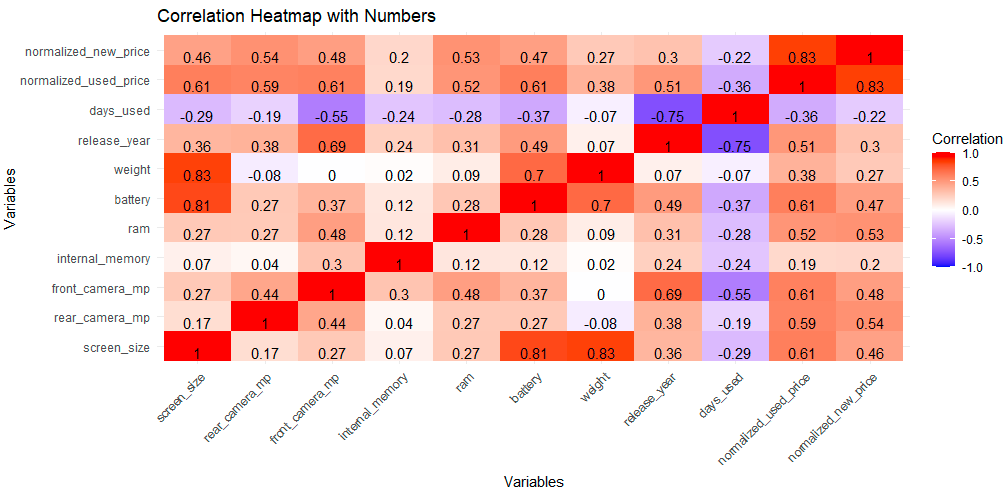
****

Fig:19

In the process of predictor selectionc The correlation coefficients from fig:19 provides insights into the strength and direction of linear relationships. Here are the key correlations observed:

* Screen size exhibits a positive correlation of 0.61 with Used\_price, indicating that larger screen sizes are associated with higher used prices.
* Rear\_camera\_mp and front\_camera\_mp both show positive correlations of 0.59 and 0.61, respectively, with Used\_price. This suggests that phones with higher megapixels in their rear and front cameras tend to have higher used prices.
* Battery also displays a positive correlation of 0.61 with Used\_price, implying that phones with larger battery capacities are associated with higher used prices.
* Release\_year exhibits a positive correlation of 0.51 with Used\_price, suggesting that newer phone models tend to have higher used prices.
* Days\_used shows a negative correlation of -0.36 with Used\_price, indicating a decrease in used prices as the number of days a phone has been used increases.
* Normalized\_new\_price demonstrates a strong positive correlation of 0.83 with Used\_price, implying that the original price of a phone strongly in fluences its used price.

|  |
| --- |
| Analysis of Variance Table |
|  |
| Response: normalized\_used\_price |
| Df Sum Sq Mean Sq F value Pr(>F) |
| Android 1 84.96 84.961 429.04 < 2.2e-16 \*\*\* |
| iOS 1 57.56 57.556 290.65 < 2.2e-16 \*\*\* |
| Others 1 32.35 32.353 163.38 < 2.2e-16 \*\*\* |
| X4g 1 275.21 275.213 1389.79 < 2.2e-16 \*\*\* |
| X5g 1 64.70 64.696 326.71 < 2.2e-16 \*\*\* |
| Residuals 3448 682.79 0.198 |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |

* The Analysis of Variance (ANOVA) was conducted to assess the impact of binary predictors on the normalized\_used\_price of mobile phones.
* The ANOVA results indicate a highly significant impact of each binary predictor (Android, iOS, Others, X4g, X5g) on the normalized\_used\_price.
* The low p-values (all < 0.001) suggest that each predictor significantly contributes to explaining the variance in the normalized\_used\_price.
* The F-statistic values are substantial, reflecting the strength of the relationship between each predictor and the response variable.

These findings affirm the importance of considering the binary features when predicting the normalized\_used\_price of mobile phones.

|  |
| --- |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) -5.452e+01 7.492e+00 -7.277 4.22e-13 \*\*\* |
| screen\_size 2.617e-02 2.740e-03 9.552 < 2e-16 \*\*\* |
| X4g 4.438e-02 1.279e-02 3.470 0.000527 \*\*\* |
| X5g -5.431e-02 2.540e-02 -2.138 0.032587 \* |
| rear\_camera\_mp 2.117e-02 1.182e-03 17.905 < 2e-16 \*\*\* |
| front\_camera\_mp 1.378e-02 9.117e-04 15.114 < 2e-16 \*\*\* |
| ram 2.191e-02 4.203e-03 5.213 1.97e-07 \*\*\* |
| battery -1.367e-05 6.117e-06 -2.234 0.025559 \* |
| weight 8.944e-04 1.062e-04 8.420 < 2e-16 \*\*\* |
| release\_year 2.764e-02 3.711e-03 7.448 1.19e-13 \*\*\* |
| days\_used 4.835e-05 2.524e-05 1.916 0.055503 . |
| normalized\_new\_price 4.276e-01 9.325e-03 45.858 < 2e-16 \*\*\* |
| iOS -7.554e-02 4.062e-02 -1.860 0.062989 . |
| Others -4.620e-02 2.471e-02 -1.870 0.061560 . |

A stepwise linear regression approach was employed for feature selection. This method systematically evaluates the inclusion or exclusion of variables based on statistical criteria, resulting in a model with a subset of features that maximizes predictive accuracy.

|  |
| --- |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) -5.354e+01 7.495e+00 -7.143 1.11e-12 \*\*\* |
| screen\_size 2.680e-02 2.678e-03 10.010 < 2e-16 \*\*\* |
| X4g 4.469e-02 1.279e-02 3.493 0.000483 \*\*\* |
| X5g -5.560e-02 2.539e-02 -2.190 0.028589 \* |
| rear\_camera\_mp 2.125e-02 1.180e-03 18.002 < 2e-16 \*\*\* |
| front\_camera\_mp 1.380e-02 9.148e-04 15.085 < 2e-16 \*\*\* |
| ram 2.249e-02 4.169e-03 5.395 7.33e-08 \*\*\* |
| battery -1.362e-05 6.129e-06 -2.222 0.026341 \* |
| weight 8.761e-04 1.049e-04 8.353 < 2e-16 \*\*\* |
| release\_year 2.715e-02 3.712e-03 7.314 3.20e-13 \*\*\* |
| days\_used 4.924e-05 2.523e-05 1.952 0.051072 . |
| normalized\_new\_price 4.275e-01 9.345e-03 45.749 < 2e-16 \*\*\* |
| iOS -7.516e-02 4.063e-02 -1.850 0.064398 . |
| Others -3.811e-02 2.498e-02 -1.525 0.127233 |

the identified features—**screen size, 4G , 5G , camera specifications, RAM, battery, weight, release year, days used, operating systems (Android, iOS, Others)**—can be considered as robust predictors for estimating used phone prices. This well-rounded selection of features not only reflects the linear relationships observed through correlation analysis but also acknowledges the impact of binary predictors.

**Data Partition:**

the dataset (Phone\_df) has been systematically partitioned into distinct subsets.

* Training Set:

Size: 2073 observations

Purpose: The majority (60%) of the dataset has been allocated to the training set. This subset will be employed for the training of our predictive models.

* Testing Set:

Size: 691 observations

Purpose: Approximately 20% of the data has been reserved for the testing set. This subset will be utilized to assess the generalizability and accuracy of our models.

* Holdout Set:

Size: 690 observations

Purpose: The remaining 20% of the data forms the holdout set. This independent subset will serve as a final evaluation metric, providing an unbiased measure of model performance on previously unseen data.

**Predictive modeling for used phone prices:**

we employed machine learning models to predict the used prices of phones in our dataset. Leveraging various predictors, including features such as screen size, network capabilities (4G and 5G), camera specifications (rear and front), RAM, battery capacity, weight, release year, usage duration, and operating system preferences (Android, iOS, and Others), we aimed to develop accurate predictive models for estimating the used prices of phones.

The training set was used to train the models, while the holdout set served as unseen data for model evaluation.

**Model Development:**

Two machine learning models were constructed to predict the used prices of phones:

Multiple Linear Regression Model: This model utilized a linear combination of predictor variables to estimate the used prices. It assumed a linear relationship between the predictors and the target variable.

Regression Tree Model: Employing a decision tree-based approach, this model partitioned the feature space into segments to predict the used prices.

**Model Evaluation:**

After training the models on the training data, we evaluated their performance on the holdout set using several evaluation metrics:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAE | RMSE | R-squared |
| Multiple linear regression model | 0.224 | 0.308 | 0.745 |
| Regression tree model | 0.255 | 0.339 | 0.67 |

Mean Absolute Error (MAE): The average absolute difference between the predicted and actual used prices.

Root Mean Squared Error (RMSE): The square root of the average squared difference between the predicted and actual used prices.

R-squared (R²): A measure of how well the model explains the variance in the used prices.

The multiple linear regression model outperformed the regression tree model across all evaluation metrics. It exhibited lower MAE and RMSE values, indicating better accuracy in predicting the used prices. Additionally, the higher R-squared value of 0.745 suggests that the multiple linear regression model explains a larger proportion of the variance in the used prices compared to the regression tree model (R-squared of 0.670).

Therefore, based on these metrics, the multiple linear regression model appears to be more effective in predicting the used prices of phones in our dataset.

**Conclusion:**

**Recommendations for Nokia:**

Product Development: Nokia can enhance its competitive position by introducing new models to the market. This strategy can capitalize on consumer demand for innovative features, ensuring Nokia remains at the forefront of technological advancements.

Camera Specifications: Improving camera specifications can further distinguish Nokia's devices in the market. With a median front camera resolution of 5 MP and rear camera resolution of 12-13 MP, Nokia can strive to match or surpass competitors such as Huawei and Samsung in camera quality. Enhanced camera features can appeal to photography enthusiasts and consumers seeking superior imaging capabilities in their smartphones.

Tablets: While Nokia exhibits a lower count of tablets compared to competitors such as LG, Samsung, and ZTE, there is potential for Nokia to expand its tablet offerings. Investing in research and development to create innovative tablet designs and features can diversify Nokia's product lineup and capture a larger share of the tablet market

Price Drop Management: Given Nokia's moderate revenue drop percentage, implementing strategies to mitigate price drops could enhance its competitive edge. These strategies may include product differentiation, value-added features, and effective marketing campaigns to maintain perceived value.

Enhanced Product Lifecycle Management: Analyzing the product life cycle highlighted certain periods of reduced device releases for Nokia. To counteract potential market downturns, Nokia could focus on innovation, product diversification, and strategic partnerships to ensure consistent product offerings throughout the year.

Operating System Transition: Nokia's transition to Android OS post-2016 aligns with prevailing market trends. Continued emphasis on Android-based devices can capitalize on consumer preferences and foster brand loyalty. Additionally, Nokia should monitor OS trends and adapt its product portfolio accordingly to remain competitive**.**

**Predictive Model selection:**

Evaluation metrics revealed that the multiple linear regression model outperformed the regression tree model in terms of accuracy and explanatory power, suggesting its suitability for predicting used phone prices

**Executive Summary**

This project explores the intricate relationships between various features of used mobile phones and their market prices. By analyzing data on features like screen size, camera quality, memory, and more, the study aims to identify the key factors that influence the pricing of used phones. Through meticulous exploratory data analysis, the project also addresses potential issues such as multicollinearity and outliers which could affect the accuracy of predictive models.

The primary objective is to provide Nokia with actionable insights by comparing features of used Nokia devices against competitors' devices. The analysis focuses on identifying trends and patterns that influence pricing and sales, thus enabling Nokia to make data-driven decisions to enhance market competitiveness and customer satisfaction.

The dataset comprises 3,454 observations and 15 variables related to phone specifications and their market prices, both new and used. Initial data cleaning involved handling 202 missing values by imputing them based on similar features within the same brand. The exploration phase highlighted the importance of certain features and their relationship with phone prices, setting the stage for deeper analysis.

The methodology of this project on used mobile phone pricing encompasses several critical phases, starting with Data Understanding and Cleaning, where the structure and quality of the data are assessed, followed by meticulous cleaning and handling of missing data. The next phase, Exploratory Data Analysis (EDA), involves exploring relationships, trends, and patterns among the variables using statistical graphics and other visualization methods. For Data Preparation for Modeling, the data is transformed to suit predictive modeling needs, which includes creating binary variables for categorical features and introducing new variables such as 'device\_type' and 'thickness' for a more detailed analysis. The subsequent stages involve Predictive Modeling where regression and classification models are built to forecast the prices of used mobile phones based on their features, and Model Evaluation and Analysis to determine the effectiveness of these models in making accurate price predictions.

Key findings from the study highlight the significant impact of phone features like screen size, camera quality, internal memory, and connectivity options (4G, 5G) on the pricing of used phones. The regression models demonstrate that newer phones equipped with advanced features tend to maintain higher market values and assist effectively in classifying phones into various price categories. A comparative analysis also revealed that while Nokia’s used devices hold a competitive edge in specific features, they fall behind in front camera quality and battery life compared to industry leaders like Apple and Samsung.

Based on the analysis, several conclusions and recommendations are provided to Nokia to enhance its market offerings. These include improving camera resolutions and battery capacity to appeal more to tech-savvy consumers and implementing targeted marketing strategies to highlight these improvements, which could boost sales and market presence. Additionally, future work could involve further analysis with a larger dataset covering more geographical regions to gain deeper insights and refining the predictive models by incorporating more features and utilizing advanced machine learning techniques to enhance prediction accuracy. These strategies are aimed at bolstering Nokia's competitive edge and driving more informed decision-making in a dynamic market environment.

**Reference**

[1] Tablet & Smartphone Resolutions and Screen Sizes List. (n.d.). Binvisions. Retrieved from <https://www.binvisions.com/articles/tablet-smartphone-resolutions-screen-size-list/>

Mortgage Payback Analytics

By Komal Singh

George Herbert Walker School of Business and Technology, Webster University

CSDA 6010: Data Analytics Practicum

Professor: Ali Ovlia

**Table of Contents**

**Introduction ……………………………………………………………………………………………60**

**Business Goal** …………………………………………………………………………………………..60

**Dataset ………………………………………………………………………………………………….61**

Attribute Definition ……………………………………………………………………………..61

Numerical Variables …………………………………………………………………………….62

Categorical Variables ……………………………………………………………………………62

**Data Preparation for Mortgage Risk Analysis ……………………………………………………….63**

Adjusting Mortgage data ………………………………………………………………………...63

**Data Cleaning …………………………………………………………………………………………..63**

Handling Missing Values ………………………………………………………………………..63

Exclusion of Specific Mortgages ………………………………………………………………..64

**Dimension Reduction and Variable Enhancement …………………………………………………..65**

Exclusion of Redundant data ……………………………………………………………………65

New Variable Introduction ……………………………………………………………………...65

Removal of Redundant variables phase 2 ……………………………………………………….66

Analysis of Remaining Variables ……………………………………………………………….66

Final Structure of Data ………………………………………………………………………….67

**Segmenting Mortgage Data …………………………………………………………………………...67**

Data frame for In-Progress Mortgages ………………………………………………………….67

Data frame for Completed Mortgages …………………………………………………………..68

Encoding of Status Time Variable ……………………………………………………………...68

**Attribute Analysis ……………………………………………………………………………………..68**

Distribution of Loan Age ……………………………………………………………………….68

Distribution of Loans by Status Time …………………………………………………………..69

Interest Rate and Loan Age Correlation ………………………………………………………...71

FICO Score Analysis ……………………………………………………………………………71

**Predictor Selection ……………………………………………………………………………………..73**

Categorical variables with Target variable ……………………………………………………...73

Numerical variables with target variable ………………………………………………………..74

Final Predictor Variables ………………………………………………………………………..75

**Data Partitioning ………………………………………………………………………………………76**

Training, Testing, and Holdout Sets …………………………………………………………….76

**Predictive Modeling ……………………………………………………………………………………76**

**Exploration of Cost Matrix for Prediction Errors …………………………………………………...77**

**Risk Flagging for In-Progress Mortgages …………………………………………………………….78**

**Executive Summary ……………………………………………………………………………………79**

**Introduction:**

The data set mortgage is in panel form and reports origination and performance observations for 50,000 residential U.S. mortgage borrowers over 60 periods. The periods have been deidentified. As in the real world, loans may originate before the start of the observation period (this is an issue where loans are transferred between banks and investors as in securitization). The loan observations may thus be censored as the loans mature or borrowers refinance. The data set is a randomized selection of mortgage-loan-level data collected from the portfolios underlying U.S. residential mortgage-backed securities (RMBS) securitization portfolios and provided by International Financial Research ([www.internationalfinancialresearch.org](http://www.internationalfinancialresearch.org)).

**Business Goal:**

Our objective is to leverage data analytics and machine learning techniques to accurately classify customers applying for mortgages as 'good' or 'bad' based on their creditworthiness and financial risk profile. Additionally, we aim to implement proactive measures to identify existing borrowers who may be at risk of defaulting on their loan payments, enabling timely intervention to mitigate potential financial losses.

**Dataset:** Mortgage dataset consists of 622,489 observations and 23 variables

|  |
| --- |
| **> str(Mortgage)** |
| spc\_tbl\_ [622,489 × 23] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame) |
| $ id : num [1:622489] 1 1 1 1 1 1 1 1 1 1 ... |
| $ time : num [1:622489] 25 26 27 28 29 30 31 32 33 34 ... |
| $ orig\_time : num [1:622489] -7 -7 -7 -7 -7 -7 -7 -7 -7 -7 ... |
| $ first\_time : num [1:622489] 25 25 25 25 25 25 25 25 25 25 ... |
| $ mat\_time : num [1:622489] 113 113 113 113 113 113 113 113 113 113 ... |
| $ balance\_time : num [1:622489] 41303 41062 40804 40484 40367 ... |
| $ LTV\_time : num [1:622489] 24.5 24.5 24.6 24.7 24.9 ... |
| $ interest\_rate\_time : num [1:622489] 9.2 9.2 9.2 9.2 9.2 9.2 9.2 9.2 9.2 9.2 ... |
| $ hpi\_time : num [1:622489] 226 225 222 220 217 ... |
| $ gdp\_time : num [1:622489] 2.9 2.15 2.36 1.23 1.69 ... |
| $ uer\_time : num [1:622489] 4.7 4.7 4.4 4.6 4.5 4.7 4.7 5 5 5.8 ... |
| $ REtype\_CO\_orig\_time : num [1:622489] 0 0 0 0 0 0 0 0 0 0 ... |
| $ REtype\_PU\_orig\_time : num [1:622489] 0 0 0 0 0 0 0 0 0 0 ... |
| $ REtype\_SF\_orig\_time : num [1:622489] 1 1 1 1 1 1 1 1 1 1 ... |
| $ investor\_orig\_time : num [1:622489] 0 0 0 0 0 0 0 0 0 0 ... |
| $ balance\_orig\_time : num [1:622489] 45000 45000 45000 45000 45000 45000 45000 45000 45000 45000 ... |
| $ FICO\_orig\_time : num [1:622489] 715 715 715 715 715 715 715 715 715 715 ... |
| $ LTV\_orig\_time : num [1:622489] 69.4 69.4 69.4 69.4 69.4 69.4 69.4 69.4 69.4 69.4 ... |
| $ Interest\_Rate\_orig\_time: num [1:622489] 9.2 9.2 9.2 9.2 9.2 9.2 9.2 9.2 9.2 9.2 ... |
| $ hpi\_orig\_time : num [1:622489] 87 87 87 87 87 ... |
| $ default\_time : num [1:622489] 0 0 0 0 0 0 0 0 0 0 ... |
| $ payoff\_time : num [1:622489] 0 0 0 0 0 0 0 0 0 0 ... |
| $ status\_time : num [1:622489] 0 0 0 0 0 0 0 0 0 0 ... |

**Attribute Definition :**

* id: Borrower ID - An identifier for each individual borrower.
* time: Time stamp of observation - Time period when the observation was made.
* orig\_time: Time stamp for origination - Time period when the mortgage loan originated.
* first\_time: Time stamp for first observation - Time period of the first observation for the mortgage loan.
* mat\_time: Time stamp for maturity - Time period when the mortgage loan matures.
* balance\_time: Outstanding balance at observation time - The remaining amount of principal on the loan at the time of observation.
* LTV\_time: Loan-to-value ratio at observation time, in % - The ratio of the mortgage loan amount to the appraised value of the property, expressed as a percentage.
* interest\_rate\_time: Interest rate at observation time, in % - The annual interest rate on the mortgage loan at the time of observation.
* hpi\_time: House price index at observation time, base year = 100 - An index measuring the movement of single-family house prices at the time of observation, with a base year of 100.
* gdp\_time: Gross domestic product (GDP) growth at observation time, in % - The percentage change in GDP at the time of observation.
* uer\_time: Unemployment rate at observation time, in % - The percentage of unemployed individuals in the labor force at the time of observation.
* REtype\_CO\_orig\_time: Real estate type condominium - 1 if the real estate type is condominium at origination time, otherwise 0.
* REtype\_PU\_orig\_time: Real estate type planned urban development - 1 if the real estate type is planned urban development at origination time, otherwise 0.
* REtype\_SF\_orig\_time: Single-family home - 1 if the real estate type is a single-family home at origination time, otherwise 0.
* investor\_orig\_time: Investor borrower - 1 if the borrower is an investor at origination time, otherwise 0.
* balance\_orig\_time: Outstanding balance at origination time - The principal amount of the loan at the time of origination.
* FICO\_orig\_time: FICO score at origination time, in % - The borrower's FICO credit score at the time of origination.
* LTV\_orig\_time: Loan-to-value ratio at origination time, in % - The ratio of the mortgage loan amount to the appraised value of the property at the time of origination, expressed as a percentage.
* Interest\_Rate\_orig\_time: Interest rate at origination time, in % - The annual interest rate on the mortgage loan at the time of origination.
* hpi\_orig\_time: House price index at origination time, base year = 100 - An index measuring the movement of single-family house prices at the time of origination, with a base year of 100.
* default\_time: Default observation at observation time - 1 if the loan is defaulted at the time of observation, otherwise 0.
* payoff\_time: Payoff observation at observation time - 1 if the loan is paid off at the time of observation, otherwise 0.
* status\_time: Default (1), payoff (2), and nondefault/nonpayoff (0) observation at observation time - A categorical variable indicating the status of the loan at the time of observation (default, payoff, or neither).

**Numerical Variables:**

|  |  |  |  |
| --- | --- | --- | --- |
| Numerical variables | | | |
| id | first\_time | LTV\_time | gdp\_time |
| time | mat\_time | interest\_rate\_time | uer\_time |
| orig\_time | balance\_time | hpi\_time | balance\_orig\_time |
| FICO\_orig\_time | LTV\_orig\_time | Interest\_Rate\_orig\_time | hpi\_orig\_time |

**Categorical Variables:**

|  |  |
| --- | --- |
| Categorical variables | |
| REtype\_CO\_orig\_time | 1 if the real estate type is condominium at origination time, otherwise 0 |
| REtype\_PU\_orig\_time | 1 if the real estate type is planned urban development at origination time, otherwise 0. |
| REtype\_SF\_orig\_time | 1 if the real estate type is a single-family home at origination time, otherwise 0. |
| investor\_orig\_time | 1 if the borrower is an investor at origination time, otherwise 0. |
| default\_time | 1 if the loan is defaulted at the time of observation, otherwise 0. |
| payoff\_time | 1 if the loan is paid off at the time of observation, otherwise 0. |
| status\_time | Default (1), payoff (2), and nondefault/nonpayoff (0) |

**Data Preparation for Mortgage Risk Analysis:**

In alignment with our business objective which is to accurately classify customers applying for mortgages as 'good' or 'bad' based on their creditworthiness and financial risk profile, the target variable selected for classification purposes is 'status\_time' which provides valuable information about the current state of the mortgage loans at the time of observation.

**Adjusting Mortgage Data:**

To prepare the mortgage dataset for analysis, several adjustments were made. Initially, a new column was introduced to identify missed payments within each ID group. This was achieved by comparing the balance at each time point with the previous balance and flagging instances where payment was missed. Additionally, the count of missed payments was calculated for each ID group.

Subsequently, a new data frame was created to contain only the last record for each ID. This step ensured that each ID was represented by a single record, containing the final status of the mortgage (default, paid off, or in progress).

The dataset now consists of 50,000 records and 25 variables.

**Data Cleaning:**

Missing Values:

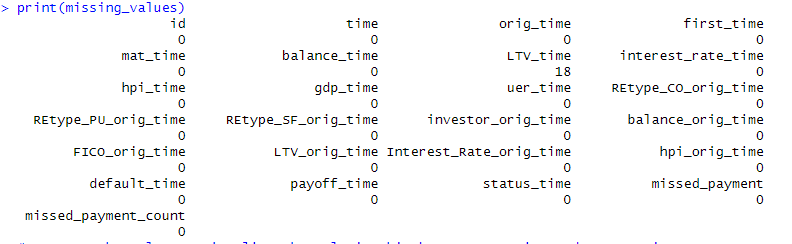


Figure : 1

18 missing values for variable LTV\_time were identified within the dataset As seen in figure:1

To ensure the integrity and reliability of our analysis, a decision was made to remove observations containing missing values. This approach was deemed appropriate given the relatively small proportion of missing data, amounting to only 18 observations out of a total dataset size of 50,000. By removing these observations, we aimed to maintain the completeness of our dataset and avoid potential biases that may arise from imputing missing values.

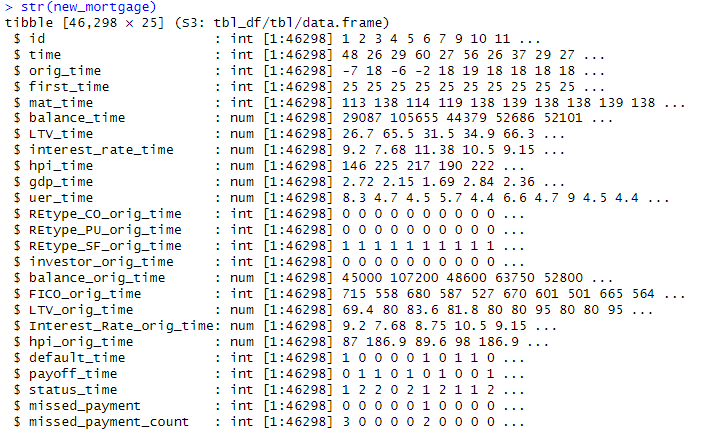
Exclusion of Specific Mortgages:

We identified a subset of mortgage records that could potentially skew the analysis due to their unique characteristics in the dataset.

**ID Count:** Mortgages represented by IDs that appear only once in the dataset. These entries are indicative of either very new or atypically short-lived mortgage records, which do not provide sufficient data points for historical comparison or pattern recognition.

**Status Time Condition:** Mortgages where the status\_time variable is set to 1 (default) or 2 (payoff). The rationale for focusing on these status times is to eliminate outlier records that could distort the predictive modeling process—particularly those that might incorrectly suggest immediate default or payoff behaviors that are not representative of typical customer profiles.

Records meeting the above conditions were removed from the primary dataset.



The Primary dataset comprises 46,298 records and has 25 variables.

**Dimension Reduction and Variable Enhancing:**

**Exclusion of Redundant data**: In the process of preparing our dataset for predictive analytics, we encountered an opportunity for dimension reduction by removing redundant information. Dimension reduction is a crucial step in data preprocessing which involves simplifying the dataset by eliminating irrelevant, redundant, or highly correlated features.

Exclusion of "default\_time" and "payoff\_time" Columns:

Excluding the "default\_time" and "payoff\_time" columns. This decision was driven by the redundancy of information present in these columns, which is already captured by the "status\_time" column.

Both the "default\_time" and "payoff\_time" columns provide information about the status of the mortgage loan at the observation time, which is essentially captured by the broader "status\_time" column. Including these columns would introduce redundancy in our dataset.

Exclusion of missed\_payment:

The decision to exclude the 'missed\_payment' column from our analysis was made following its initial utilization to obtain counts on missed payments within the mortgage dataset.

Exclusion of Interest rate at origination time:

We identified that the Interest\_Rate\_orig\_time column, which represents the interest rate at the time of loan origination, was redundant. This redundancy was observed because the interest rate remained unchanged between the time of origination and the time of observation across all records. therefore, excluding the interest\_rate\_time column.

**New Variable Introduction:** The enhancement of our dataset involved the creation of novel variables derived from the existing data. This process is pivotal in predictive modeling as it unveils additional dimensions of information that can be critical for understanding the dynamics of mortgage performance.

New Maturity Time: Enhancement of Maturity Time Calculation

To provide a more accurate representation of the mortgage duration by recalculating the maturity time to account for loans originated before the commencement of the observed data period. In our dataset, orig\_time is recorded such that it can start at negative values if the loan originated before the start of our observation window. This can lead to misrepresentations in calculating the actual loan duration if not adjusted properly.

The original mat\_time column indicates the scheduled end of the loan, but does not account for loans starting at negative times effectively. To address this, we introduce new\_maturity\_time which adjusts the maturity time based on when the loan originated. The new maturity time is computed as the difference between mat\_time and orig\_time.

Loan Age:

The loan\_age variable provides insight into the time elapsed since the origination of the mortgage. This measure is particularly salient as it may correlate with borrower behavior, influencing the likelihood of prepayment or default.

Time to Maturity:

Building upon the new\_maturity\_time, we derive the Time\_to\_maturity variable, which denotes the remaining duration until the loan reaches maturity from the current observation point. This variable is crucial for modeling as it directly impacts the risk assessment and valuation of the mortgage loan.

Housing Price Index Change Percentage:

The HPI\_Change\_Percentage variable captures the relative change in property values as reflected by the housing price index from the time of loan origination to the observation date. This metric is a proxy for equity changes in the collateral property and can influence borrower decisions regarding refinancing or default.

**Removal of redundant Variables phase 2:**

This process involved eliminating certain variables from the dataset while retaining the essential information needed for analysis and modeling. Below are the columns that were removed from the original dataset:

**time**

**orig\_time**

**first\_time**

**mat\_time**

**hpi\_time**

**hpi\_orig\_time**

**new\_maturity\_time**

These columns were redundant after the creation of new variables derived from them. Specifically, we generated new variables to capture essential information such as loan age, time to maturity, and percentage change in Housing Price Index (HPI).

**Analysis of Remaining variables:**

While several columns were excluded or transformed during the dimension reduction and variable enhancing process, certain key variables were retained due to their significance in mortgage performance analysis and predictive modeling.

gdp\_time and uer\_time:

These variables remain in the dataset to capture economic indicators such as GDP and unemployment rates, which are crucial for assessing the overall economic environment and its impact on mortgage loan performance.

FICO\_orig\_time:

Retaining borrowers' FICO scores at loan origination allows for a nuanced evaluation of creditworthiness, enhancing the predictive accuracy of models regarding default risk.

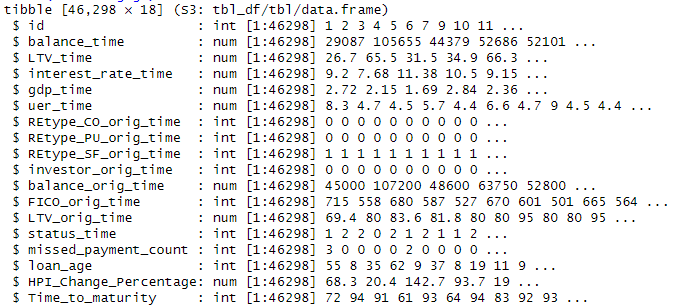
RE\_type:

The inclusion of Real Estate Type information enables a better understanding of the diverse risk profiles associated with different property types, contributing to more informed decision-making in mortgage analysis and forecasting.

LTV (Loan-to-Value ratio):

By keeping the Loan-to-Value ratio, we ensure consideration of borrowers' equity positions, which is essential for assessing the risk exposure of lenders and understanding potential default scenarios in mortgage portfolios.

**Final Structure of the data:** Initially, our dataset consisted of 25 columns, including variables such as time, orig\_time, first\_time, mat\_time, hpi\_time, hpi\_orig\_time, and new\_maturity\_time, among others. However, through careful examination and feature engineering, we identified opportunities to condense the dataset without sacrificing predictive power. As a result, we derived new variables that captured key temporal and economic insights, effectively reducing the dimensionality of our dataset to 18 variables.



**Segmenting Mortgage Data:**

We have undertaken a strategic approach to segment the mortgage dataset into two distinct data frames. This segmentation is aimed at facilitating our business objectives effectively.

We have created two separate data frames based on the status of mortgages:

a. Data frame for In-Progress Mortgages:

This data frame exclusively consists of mortgages that are currently in progress, denoted by a status code of 0. These mortgages are actively being serviced, and our primary focus here is to identify potential red flags or indicators of financial distress. By isolating these cases, we aim to implement proactive measures to mitigate risks and prevent potential defaults.

b. Data frame for Completed Mortgages:

In contrast, this data frame comprises mortgages that have either been paid off (status code 1) or defaulted (status code 2). This subset serves as the basis for training and evaluating machine learning models. By leveraging historical data from completed mortgages, we can train models to accurately predict the likelihood of default or successful repayment for new applicants.This data frame will be used for the analysis purpose.

Encoding of Status Time Variable:

To ensure clarity and consistency in our analysis, we have encoded the "status\_time" variable such that a value of 1 corresponds to default, while a value of 0 corresponds to payoff. This standardized encoding scheme facilitates seamless differentiation between default and payoff scenarios, streamlining subsequent classification and risk assessment tasks.

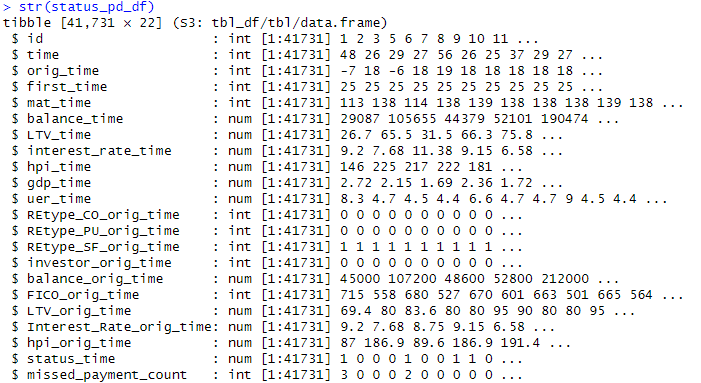


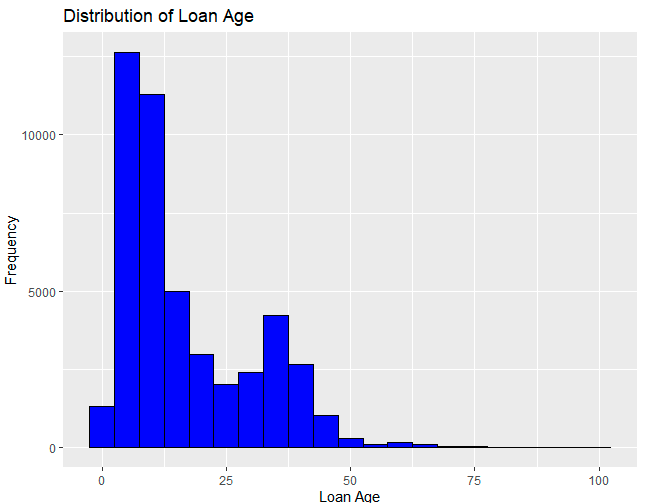
Figure:2

Figure 2 represents the final structure of the data frame after cleaning and preprocessing the data. the data frame consists of 41,731 observations and 22 variables.

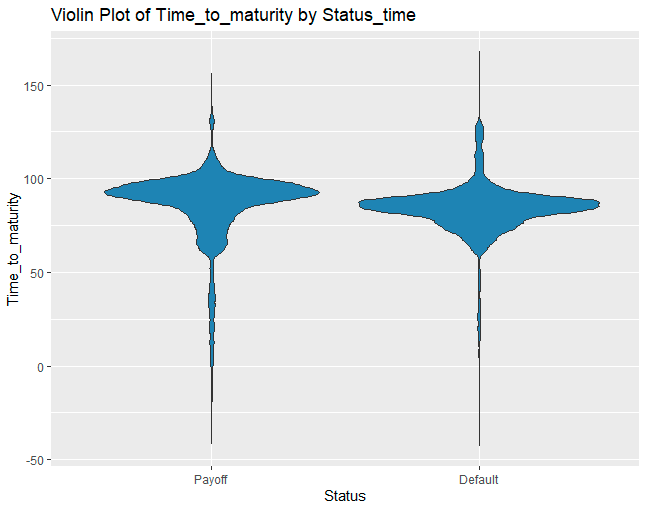
**Attribute Analysis:**

Distribution of loan age:

The histogram shows the distribution of loan ages with most loans being relatively new and a sharp decline in frequency as loan age increases. This indicates that fewer loans reach maturity, with the majority being concentrated in the lower age bracket.



Distribution of loans by status\_time by Time to Maturity:



The violin plot illustrates the distribution of time to loan maturity, categorized by two statuses: payoff and default. Both distributions show a wide range of times to maturity but with different shapes and spreads. The payoff status has a more pointed distribution at the top, while the default status appears more uniformly distributed. This suggests variability in how long loans exist before being paid off, with defaults occurring across a more consistent time range.

Distribution of Balance amount for different Real estate type:

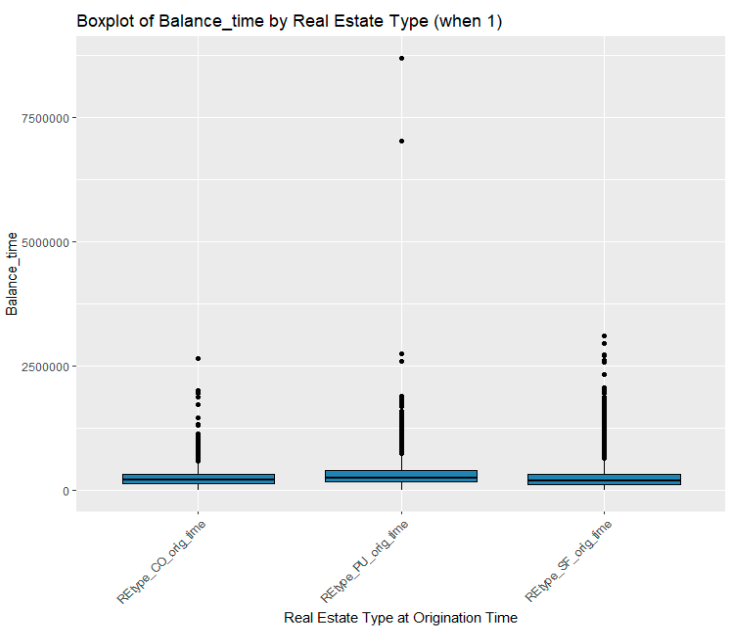
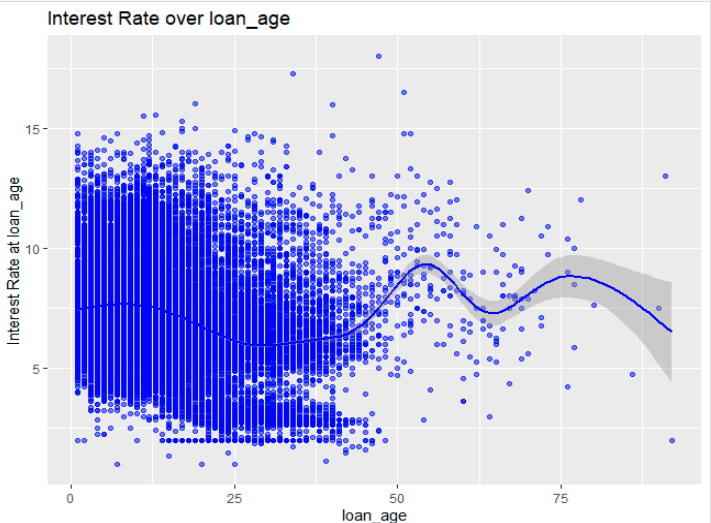


Figure: 6

This graph in Figure 6 is a boxplot displaying the distribution of outstanding balances (balance\_time) for three different types of real estate at the time of observation. The types are condominiums (REtype\_CO\_orig\_time), planned urban developments (REtype\_PU\_orig\_time), and single-family homes (REtype\_SF\_orig\_time).

* Condominiums: The median balance is relatively low compared to the other types, with a compact IQR, suggesting that the outstanding balances for condominiums are generally lower and closer together.
* Planned Urban Developments: The median balance is higher than that for condominiums, and the IQR is wider, indicating more variation in outstanding balances.
* Single-Family Homes: The median balance is similar to that of planned urban developments, but the range of outliers is quite extensive, suggesting that while many single-family home loans have balances similar to those of planned urban developments, there are also a number of them with significantly higher outstanding balances.

Distribution of interest rate over Loan age:



This scatter plot with a trend line shows the relationship between loan age and interest rate. Initially, interest rates are quite varied, but they seem to decrease and stabilize as loans age. The shaded area suggests increased uncertainty in interest rates for older loans

Box plot of FICO score by status time:

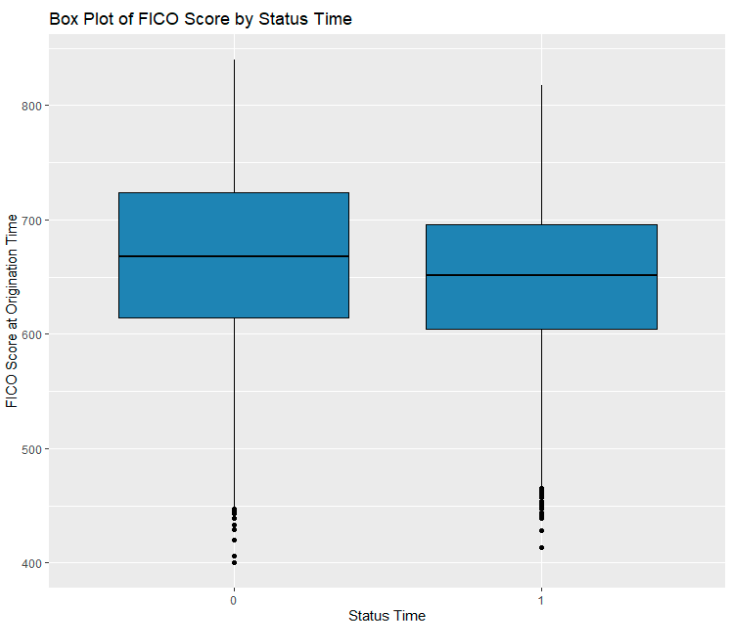


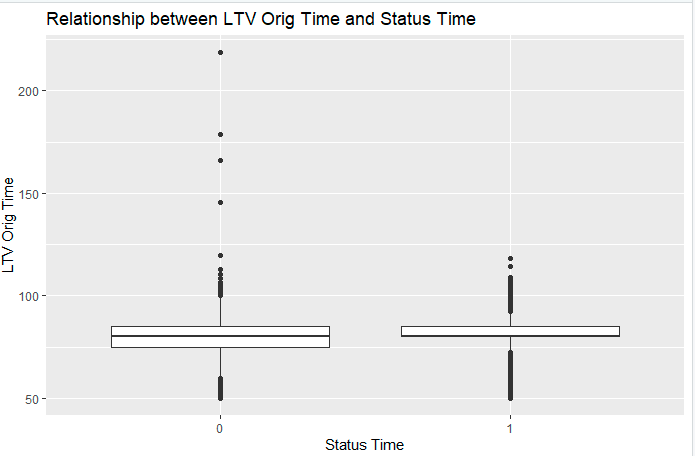
Figure: 9

This graph in Figure 9 is a boxplot representing the distribution of borrowers' FICO credit scores at the time of loan origination, categorized by the status of the loan at the time of observation—either default (1) or payoff (0).

The median FICO score for both categories is above 600, with the median for loans that are paid off being slightly higher. The box representing loans that have been paid off (status\_time 0) has a narrower interquartile range (IQR), indicating that the FICO scores for this group are more tightly grouped than for those that defaulted.

The box representing loans that defaulted (status\_time 1) has a lower IQR, but not by a substantial margin. This suggests that borrowers who default have a wider range of FICO scores but trend towards the lower end. There are outliers in both categories, with several particularly low FICO scores associated with defaults. Outliers on the payoff side suggest that there are some borrowers with low FICO scores who still manage to pay off their loans.

Relationship between LTV at origination time and status of the loan:



The LTV\_orig\_time is a continuous variable and is represented on the y-axis. The status\_time is a categorical variable with two categories and is shown on the x-axis, where "0" represents loans that have been paid off, and "1" represents loans that have defaulted.

The graph suggests that there is a relationship between the LTV ratio at the origination time and the likelihood of default. A higher LTV ratio at the origination time appears to be associated with a higher likelihood of default.

**Predictor selection:**

**Categorical variables with Target variable :**

Chi square test was performed between all the categorical variables and target variable status\_time to check their significance.

|  |
| --- |
| Predictor Chi2\_Statistic P\_Value |
| X-squared REtype\_CO\_orig\_time 1.5893478 2.074193e-01 |
| X-squared1 REtype\_PU\_orig\_time 8.9589084 2.761198e-03 |
| X-squared2 REtype\_SF\_orig\_time 0.6225219 4.301117e-01 |
| X-squared3 investor\_orig\_time 53.5234120 2.555294e-13 |

REtype\_CO\_orig\_time: The Chi-square statistic is 1.5893 with a p-value of 0.207. This indicates no statistically significant association between REtype\_CO\_orig\_time and status\_time at the 0.05 significance level.

REtype\_PU\_orig\_time: The Chi-square statistic is 8.9589 with a p-value of 0.0028. This p-value is less than 0.05, suggesting a statistically significant association between REtype\_PU\_orig\_time and status\_time.

REtype\_SF\_orig\_time: The Chi-square statistic is 0.6225 with a p-value of 0.430. This implies no statistically significant relationship between REtype\_SF\_orig\_time and status\_time at the 0.05 significance level.

Investor\_orig\_time: The Chi-square statistic is 53.5234 with a highly significant p-value of approximately 2.56e-13. This p-value is much less than 0.05, indicating a strong statistically significant association between investor\_orig\_time and status\_time.

**Numerical variables with target variable:**

Analysis of Variance (ANOVA) was employed to test for statistically significant differences in the means of various numerical predictors across the different categories of the target variable, status\_time.

Specifically, we performed a one-way ANOVA for each numerical predictor against status\_time to determine if the average values of these numerical variables are different for loans that default compared to those that do not. A significant result from ANOVA indicates that the numerical predictor may play a role in the loan's outcome and should be further investigated.

|  |
| --- |
| Predictor F\_Value P\_Value |
| 1 loan\_age 140.428759 2.445638e-32 |
| 2 Time\_to\_maturity 136.229994 2.010552e-31 |
| 3 balance\_time 47.429152 5.791252e-12 |
| 4 LTV\_time 12134.498742 0.000000e+00 |
| 5 interest\_rate\_time 192.214126 1.334213e-43 |
| 6 HPI\_Change\_Percentage 6812.551733 0.000000e+00 |
| 7 gdp\_time 4764.684730 0.000000e+00 |
| 8 uer\_time 4528.322372 0.000000e+00 |
| 9 balance\_orig\_time 0.243162 6.219345e-01 |
| 10 FICO\_orig\_time 523.016451 5.582246e-115 |
| 11 LTV\_orig\_time 306.995993 1.831911e-68 |
| 12 missed\_payment\_count 970.816604 1.770240e-210 |

loan\_age and Time\_to\_maturity:

These two predictors have high F-values and very low p-values, indicating a statistically significant relationship with the target variable (status\_time).

Specifically, the low p-values (close to zero) suggest strong evidence against the null hypothesis, indicating that both loan\_age and Time\_to\_maturity are likely relevant predictors of status\_time.

balance\_time, LTV\_time, interest\_rate\_time, HPI\_Change\_Percentage, gdp\_time, and uer\_time:

These predictors also exhibit high F-values and extremely low p-values, suggesting highly significant relationships with status\_time.

The p-values are effectively zero, indicating strong evidence that these variables are important predictors of the target variable.

balance\_orig\_time:

The F-value for this predictor is relatively low, and the p-value is high (0.6219).

This suggests that there is insufficient evidence to conclude a significant relationship between balance\_orig\_time and status\_time.

FICO\_orig\_time and LTV\_orig\_time:

These predictors have high F-values and extremely low p-values, indicating a highly significant relationship with status\_time.

The p-values are close to zero, providing strong evidence that both FICO\_orig\_time and LTV\_orig\_time are important predictors of the target variable.

missed\_payment\_count:

This predictor exhibits a high F-value and a very low p-value, indicating a statistically significant relationship with status\_time.

The low p-value suggests strong evidence against the null hypothesis, indicating that missed\_payment\_count is likely a relevant predictor of status\_time.

**Final List of Predictors:**

|  |  |  |
| --- | --- | --- |
| # | Varaible type | Variables |
| 1 | Categorical | REtype\_PU\_orig\_time |
| 2 | Categorical | Investor\_orig\_time |
| 3 | Numerical | missed\_payment\_count |
| 4 | Numerical | LTV\_orig\_time |
| 5 | Numerical | FICO\_orig\_time |
| 6 | Numerical | HPI\_Change\_Percentage |
| 7 | Numerical | gdp\_time |
| 8 | Numerical | uer\_time |
| 9 | Numerical | Loan\_age |
| 10 | Numerical | Time\_to\_maturity |

**Data Partition :**

In the development and validation of our predictive model, we recognized the importance of accurately assessing its performance and generalizability to unseen data. To achieve this, we partitioned our dataset into three distinct subsets: a training set, a testing set, and a holdout set.

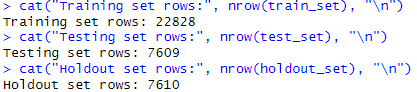
The training set is utilized to build and train the model, allowing the algorithm to learn the underlying patterns within the data. The testing set is used to evaluate the model's performance and to fine-tune its parameters to optimize its accuracy. Finally, the holdout set serves as an additional layer of validation, used to test the model after it has been finalized, simulating its application on future, unseen data.

For our dataset, we adopted the following partitioning strategy:

Training set: 60% of the data,

Testing set: 20% of the data,

Holdout set: 20% of the data.

****

The resulting subsets consisted of:

Training set: 22,828 rows

Testing set: 7,609 rows

Holdout set: 7,610 rows

Utilizing In-Progress Mortgage Data for Red Flag Identification:

The data frame status\_0\_df, previously segregated to exclusively contain in-progress mortgage data, will serve a critical role in our analysis. It will be employed to identify potential red flags by leveraging the model that was trained on a comprehensive dataset consisting both paid-off and defaulted mortgages.

**Model Selection for Predicting Mortgage Outcomes:**

Classification Models for Outcome Prediction:

To predict mortgage outcomes, we considered several classification models known for their predictive power and efficiency in handling binary outcomes. Among these, the classification tree stands out for its simplicity and interpretability. Classification trees work by repeatedly splitting the data into smaller and smaller subsets based on certain criteria, making them particularly useful for understanding the decision rules that lead to a mortgage being classified as good or bad. we did not limit our exploration to classification trees alone. We also will be trying out different models, including Random Forests and Gradient Boosting Machines (GBMs),

we applied three classification models to predict the status of mortgage loans based on a set of predictor variables. The models applied are Logistic Regression, Classification Tree, and Naïve Bayes. We evaluated the performance of each model on a test data set and a holdout data set.

|  |  |  |
| --- | --- | --- |
| Model | Test data Accuracy | Holdout data Accuracy |
| Logistic regression | 0.7567354 | 0.7533509 |
| Classification Tree | 0.7715863 | 0.7613666 |
| Naïve Bayes | 0.7307136 | 0.7282523 |

Based on the accuracy results, the Classification Tree model achieved the highest accuracy on both the test and holdout data sets, followed closely by the Logistic Regression model. The Naïve Bayes model had the lowest accuracy on both data sets.

The Classification Tree model had a slight edge in accuracy on both data sets, making it the preferred choice for predicting the status of mortgage loans. The Logistic Regression model also performed well and is a solid theoretical choice for binary classification problems. The Naïve Bayes model had the lowest accuracy and may not be the best choice for this problem due to its assumption of independence among predictors.

Classification Tree Metrics:

Based on the confusion matrix, the Classification Tree model predicted 3907 loans as status 0 and 1887 loans as status 1 on the holdout data set. However, there were 716 loans that were actually status 0 but were predicted as status 1, and 1100 loans that were actually status 1 but were predicted as status 0.

**Exploration of Cost Matrix for Prediction Errors:**

In the process of evaluating our classification model, significant effort was dedicated to developing a cost matrix aimed at quantifying the financial implications of prediction errors. This involved calculating average balance\_time for both defaulters and non-defaulters within our dataset. Specifically, the cost of false positives (FP) was set as the average balance\_time for defaulters, reflecting the direct financial loss associated with incorrect default predictions. Conversely, the cost of false negatives (FN) was determined by applying an interest rate to the average balance\_time for non-defaulters, intending to capture the financial cost of missed opportunities due to underestimating the risk.

While this method provided valuable insights into the potential costs associated with each type of prediction error, the results suggested unusually high costs. This was primarily due to the elevated average balance\_time, which significantly influenced the calculated costs. Given these findings and the inherent uncertainties about the appropriateness of the methodology—particularly the assumptions made regarding the impact of balance times and the applied interest rates—it was decided not to include this cost matrix in the final report.

**Risk Flagging of In-Progress Mortgages:**

Objective:

As part of our risk management strategy, we applied our classification model to status\_0\_df, a dataset encompassing all in-progress mortgages. The goal was to identify loans that may carry a higher risk of default, thereby enabling proactive measures to mitigate potential financial losses.

Methodology:

We employed our previously validated Classification Tree model, which demonstrated the highest accuracy across test and holdout datasets, to predict the likelihood

of default for each in-progress mortgage. The model predictions were based on a comprehensive set of predictors, including original loan characteristics, payment history, and economic indicators such as GDP and unemployment rate, amongst others.

Risk Flagging Process:

The risk flagging was conducted by computing the predicted probability of default for each mortgage in the dataset. A threshold was set such that any mortgage with a probability of default above this threshold was flagged as "High Risk", and those below were flagged as "Low Risk".

Threshold Determination:

The threshold for risk flagging was set at 0.5, a standard practice for binary classification problems. This value was chosen to balance the sensitivity and specificity of our model; however, it is adjustable based on the risk appetite of our financial institution.

Results:

The application of the model resulted in the flagging of 2,561 mortgages as "High Risk" and 5,690 as "Low Risk". This indicates that approximately 31% of the current in-progress mortgages have been classified as high-risk loans.

|  |
| --- |
| High Risk Low Risk |
| 2561 5690 |

**Executive Summary:**

This project report encapsulates the rigorous analysis undertaken to predict mortgage payback capabilities using a comprehensive dataset of 50,000 U.S. residential mortgages over 60 periods. The data, sourced from U.S. residential mortgage-backed securities, provided a randomized selection representative of the wider market.

The crux of our endeavor was to leverage advanced data analytics and machine learning techniques to classify potential mortgage borrowers into 'good' or 'bad' categories, based on their creditworthiness and likelihood of default. We also sought to proactively identify existing borrowers at risk of defaulting, aiming for timely interventions to curtail financial losses.

Our analysis was underpinned by a dataset encompassing 622,489 observations across 23 variables, including borrower ID, loan balance, interest rates, and economic indicators such as GDP growth and unemployment rates. Essential data cleaning and preparation processes were executed, including handling of missing values and exclusion of anomalous mortgage records to ensure data integrity and relevance.

Key to our analytical framework was the creation of new variables, like Loan Age and Time to Maturity, and the introduction of a risk flagging mechanism for in-progress mortgages. The meticulous attribute analysis, including Chi-square tests for categorical variables and ANOVA for numerical predictors, informed the selection of the most impactful predictors.

Our predictive modeling leveraged three classification techniques—Logistic Regression, Classification Tree, and Naïve Bayes—with the Classification Tree model outperforming others in accuracy. This model was then applied to in-progress mortgages to flag potential defaults. Setting a risk threshold at 0.5, we flagged 2,561 mortgages as 'High Risk' and 5,690 as 'Low Risk', revealing that approximately 31% of the analyzed in-progress mortgages could potentially default.

The analysis conducted and the subsequent risk flagging offer a significant tool for risk management within the mortgage industry. Despite the calculated financial impact of prediction errors suggesting higher-than-expected costs, the decision to exclude these from the final report reflects our commitment to data integrity and methodological accuracy. The insights garnered from this project not only demonstrate the applicability of data analytics in financial risk assessment but also underscore the importance of continuous model validation and adjustment in predictive analytics.