PRODUCTS DUMMY DATA ANALYSIS

In this analysis I chose to focus on the regression model to predict the amount paid by the 360th day.

1. Data Pre-Processing

The dataset contained 13 attributes and 9,195 observations. Investigation of the data set led me to drop the 'Total Days' and 'Start Date' variables since all the loan required the customer to pay within 365 days. The day the customer received the loan and the total days required to pay back the loan can therefore not explain the variations in the amount of the loan paid as of the 360th day.

The data was relatively clean since it had consistent column names, no missing data and appropriately initialized variables. The 'Gender' column however included an inappropriate value- 'Not Specified'. These values were transformed into either male or female based on the total price of the product purchased by the customer since there was a slight difference in the variable across gender.

Variable specific analysis revealed that the location variable contained very many unique observations which would hinder the creation of a parsimonious model. As a result, I dropped the variable and used the region variable instead since it still captured information on the location of the customers.

2. Feature Engineering

The target variable in this study is the loan amount paid. However using this variable as is may be misleading as it implies that customers who has paid the same amount are equally desirable without taking into consideration the total amount of the purchased product. For example, consider two customers who purchased 2 different products. Customer A purchased product C for Kshs.13,775 and customer B purchase product F for Kshs.69,225. Supposing the two customers each paid 6,922.50 toward their loan. Viewing the absolute amount paid would consider both customers equally desirable. However, in the business sense, customer A is preferred to customer B since the amount paid represents 50% of customer A's purchase and only 10% of customer B's purchase.

Further, since the absolute value of the loan paid is not limited and can take on any amount, it would be possible to estimate an amount paid which is beyond the total price of a product. The latter would be erroneous since no one would be willing to pay more than the quoted price of a product in the real world.

Similarly, the variables detailing the amounts paid by the 30th and 60th days were converted into percentage terms. The values were calculated as percentages of what was supposed to have been paid by the 30th and 60th days according to the payment plans and including the deposit. Comparing the amount paid by the amount required is better than comparing the amount paid to the total price as the latter gives us information about if the customer adheres to the payment plan or not. We are therefore able to estimate the amount a customer is likely to pay based on how well they adhere to the payment plan. As for the amount paid by the 360th day I compared the value with the total amount since some customers have finished their payments meaning that if I compared the amount paid by the 360th day to the amount required by that day, I would get a value greater than 100% which is undesirable.

3. Data Exploration

a.) Products

Different products had different payment plans as detailed below:

Product	TotalPrice	DailyRate	Deposit
Product A	19250	50	1000
	20250	50	2000
	20750	50	2500
	22250	50	4000
Product B	16600	40	2000
Product C	13775	35	1000
	15600	40	1000
Product D	16600	40	2000
	20250	50	2000
Product E	53625	125	8000
	54625	125	9000
Product F	69225	165	9000

The products B and A were the most popular products purchased during the period under study. On the other hand, products E and F were the least desirable with only 5 units of product F sold. The latter may have been influenced by price since product F was the most expensive whereas product B was on the most affordable options.

The percentage of payments made by the 360th day were on average below the expected 98.6% across different product lines. Payments for product A and F recorded the worst performance.

A majority of the customers were active customers (73.3%), 9.2% were blocked while 17.4% had finished their payments. Customers of products A and B formed approximately 66% of the customers.

Product	Product A	Product B	Product C	Product D	Product E	Product F	All
LoanStatus360							
Active	0.277325	0.302773	0.064274	0.067972	0.020555	0.000326	0.733225
Blocked	0.027515	0.047308	0.008809	0.007939	0.000653	0.000109	0.092333
Finished Payment	0.036759	0.069929	0.035563	0.030125	0.001958	0.000109	0.174443
All	0.341599	0.420011	0.108646	0.106036	0.023165	0.000544	1.000000

b.) Gender

During the review period, more men purchased items as opposed to women. On average, men also purchased items worth more than women on and paid 1% more than women towards their loans. Product B was purchased most across

c.) Regions

Assumption that amount required includes deposit

Describe

Test of OLS Assumptions

What is a negative amount paid?

Hypothesis testing?

Do anova for each category

Put a scatter plot vs the model

R squred

Add visualizations

Regional variable had too many observation and it would be better to further consolidate the regions so that we can have fewer categories therby increasing our degrees of freedom and improving precision by increasing observations.

Confusion matrix