**Final Project**

**STAT311-50**

**Fall 2024**

**Ian Wallin**

**An Analysis on the Effect of Bank Customers on Generated Revenue**

**Introduction:**

A bank collected data on customers in an effort to predict the revenue collected from accounts; in total, the bank collected data on 7,420 customers. The data collection includes quantitative variables: total revenue earned from the account over a 6-month period, balance total across all accounts owned by each customer, age of the customer, and the number of years the customer has held an account. An indicator variable was collected to indicate whether a customer had received a promotional offer in the previous one-month period. Additionally, the bank provided other indicator variables for the account types a customer may have; a 1 was assigned to this variable if the associated account type has a high level of activity, and a 0 if there is little to no activity. The available account types offered by the bank are:

Checking

Debit Card

 Savings

 Credit Card

 Loan

 Mortgage

 Insurance

 Pension

 CD

 Money Market

 Non-Primary Savings.

The bank would like to determine what type of customers are most likely to generate revenue for the bank and attempt to predict the revenue potential. This will assist the bank in determining what type of customers to target, and what types of accounts to promote to maximize revenue.

**Methods:**

As standard practice, we will begin our analysis with some exploratory data analysis and see how our data is distributed to try an anticipate steps that may need to be taken when building a predictive model. From there we will start with a basic model, including only the quantitative variables: Balance Total, Customer Age, and Account Age. Doing so will provide us some initial residuals, hat values, and Cook’s Distance values that will help in eliminating any influential outliers in the data. Once any necessary exclusions are made, we will have to address the remaining variables. There are eleven account types offered an we also have to consider the variable whether a customer received an offer within the previous one-month period. We anticipate that not every variable is influential and attempting to build a higher order model if necessary, would be arduous. Our plan instead is to build an initial model that does include the twelve additional variables but does not include interactions or second-order variables. This is done as an extension of the exploratory data analysis to gain insight on these variables and how the overall model performs with their inclusion. To create a more reasonable and maintainable model, we will use Stepwise methods to find more crucial variables to be included. Once the finalized variables have been decided we will again look at model performance and can look at further tweaking or building a higher order model if necessary.

**Exploratory Data Analysis**

We’ll start our exploratory data analysis by looking at the distribution of Total Revenue by means of a Histogram and a Box and Whisker Plot shown in Figure 1 below:

A screenshot of a graph

Description automatically generated

Figure

We can immediately see that our data is heavily right-skewed with a lot of outliers. If we take a look at the Summary Statistics below in Figure 2 we can calculate what values would be considered outliers.

A screenshot of a computer

Description automatically generated

Figure

From Figure 2, we have a standard deviation of 3.21 meaning that values that fall outside three standard deviations would be considered outliers. This mean that values falling outside 9.63 would be outliers, a considerable amount of our Revenue Total value seem to fall outside this range. To avoid removing too much data unnecessarily, we’ll hold off on excluding these values until we have created an initial model and generate some values by which to test their influence. Before we build our first model, let’s look at a scatterplot of Revenue Total vs each of the quantitative variables individually show below in Figure 3:

A screenshot of a graph

Description automatically generated

Figure

Looking first at all three plots highlights why we may not want or need to remove all outliers. We can see the some point fall outside the larger clusters of data but many follow the same trend as the tighter grouped data. On the Rev\_Total vs Bal\_Total plot there appears to be a downward trend where the higher the Balance Total is, the lower the Revenue Total is, this may indicate that targeting customers who will carry lower balances may yield higher revenue. On the Rev\_Total vs AGE plot, we see what appears to be a non-linear trend with Revenue increasing as Age increases until around Age 50 when Revenue then curves downwards, meaning a higher order term may be needed in the model should this variable be influential. Finally, we look at the Rev\_Total vs AccountAge; this plot has less of a discernible trend, though there is a noticeable drop in revenue on accounts older than 16 years old approximately.

**Analysis**

**Model 1:**

First Order Model of Quantitative Variables

Where:

**Hypothesis:**

**Output:**

**A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph with red lines and numbers

Description automatically generated**

**A graph with numbers and a line

Description automatically generated**

Model 1 produces a questionable model with an value of only 0.0301 indicating the model only accounts for 3.01% of the variance. The F value of 77.8507 and the p-value of < 0.0001 means that we would reject the null hypothesis, and the model is statistically viable at predicting revenue. Looking at the Parameter Estimates table, we see that individually each variable’s p-value is < 0.0001 indicating each variable is statistically significant in predicting revenue total. We should investigate these results further to determine what can be improved.

Residual Histogram Plot

A graph with a black line

Description automatically generated

Residual Normal Quantile Plot

A graph with a line

Description automatically generated

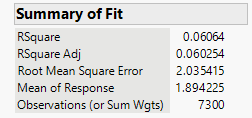
We see from the Residual distribution plot, similar to the one produced in our exploratory data analysis, that our data is very right-skewed. We will test to see if these outliers are influential on the effectiveness of our model. The value that is the largest outlier, an account with a revenue total of 94.1, has a Cook’s Distance value of only 0.035 which does not exceed the threshold of 1.0 to be considered influential. If we go back to Model 1’s output and review the Externally Studentized Residuals plot, shown below, we see there are a lot of values that fall outside the Bonferroni limit in red and indicate these values are probably influential.

Externally Studentized Residuals

A graph showing a row number

Description automatically generated

On the residual distribution plot we had a standard deviation of 3.158, meaning that values outside 9.474 are outliers. If we exclude values outside this range, we remove 120 rows of the 7,420 rows of data we have collected, leaving 7,300 rows which should still be more than sufficient to make predictions. Re-running Model 1, we see some improvements, but still a lot of room to grow:

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph of a graph showing a number of dots

Description automatically generated with medium confidence

A graph of a number of dots

Description automatically generated with medium confidence

A graph with a line

Description automatically generated

A graph with a line and numbers

Description automatically generated

There is still a lot of vertical data that deviates on the Actual vs Predicted plot though it’s less severe than the original model’s plot. Looking at Summary of Fit, our value has doubled from 0.03 to 0.06, meaning this new model accounts for 6% of the variance. We also have a much higher F value of 156.997 compared to the previous 77.85 and again a low p-value of < 0.0001. Looking at the Studentized Residual Plot and the Normal Quantile Plot, they are improved, though far from ideal. The issue seems to be that a lot of the data is near the origin on the x-axis with a high level of variance on the y-axis. From here we will try and find additional variables to improve this model. Starting by including all variables and then trying to find the necessary ones.

**Model 2:**

All variables included, no higher order terms or interactions:

Where:

**Hypothesis:**

**Output:**

**A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph of a graph showing a number of dots

Description automatically generated with medium confidence**

**A graph showing a number of dots

Description automatically generated**

We’ll go through this model in the same manner as the previous one. Starting with the we again see an improvement with a value of 0.14779, indicating the model accounts for 14.779% of the variance. This is again a little more than double the value from the previous model of 6%. The p-value is also < 0.0001 indicating we would reject the null hypothesis and the model is statistically viable to make revenue predictions. A final improvement we see from the statistic tables is the Mean Square Error has decreased from 4.143 to 3.757, not a huge improvement but still better. We do see though that the F value has decreased from 156.9975 to 106.4822. We would still consider this model an improvement over the previous one though. Looking at the Actual vs Predicted plot we see tighter clustering which is also a good indication of improvement. We believe that we can still improve upon this model; the Parameter Estimates table shows that some of the added variables have p-values > 0.05 significance level and are not statistically significant to the model. Our next step will utilize Stepwise methods to find what variables are vital for a predictive model and build from there.

Running Stepwise Fit for Rev\_Total, we started by selecting All Possible Models and looked at the models with the lowest Cp and highest R2 values, the output is shown below:

A screenshot of a computer

Description automatically generated

We see consistently that all models with the highest R2 values and lowest Cp values include seven independent variables with the most consistent variables included being: Bal\_Total, Age, AccountAge, Offer, Card, and LOAN, with the last variable differing between models. Running a Backward Stepwise Fit with a P-value threshold of 0.05 to enter and leave, the selected variables are shown below:

A screenshot of a computer

Description automatically generated

We see a similar output but AccountAge was removed. AccountAge being removed makes some sense when looking at the plots we created during our exploratory data analysis; this variable had a less discernable pattern. Running a Forward and Mixed Stepwise Fit both produce the same model as the Backward. With these tests run, let’s test the models from when All Possible Models were tried with the seventh variable being INSUR as that was consistently selected on the other Stepwise Models. We will call this Model 3 which is shown below:

**Model 3:**

**Where:**

**Hypothesis:**

**Output:**

**A screenshot of a computer

Description automatically generated** **A screenshot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

We’ll also run a model with AccountAge removed to see if this affects the model positively or negatively and call this Model 4:

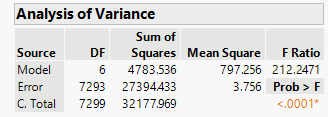
**Model 4:**

**Where:**

**Hypothesis:**

**Output:**

**A screenshot of a computer

Description automatically generated** ****

**A screenshot of a computer

Description automatically generated**

We see from both models’ outputs there is almost no noticeable difference in their viability. Both models have a P-value of < 0.0001, an value of .148 meaning both models account for 14.8% of the variance; this is the almost the same value produced in Model 2 as well with the main benefit of this model being that we have eliminated 9 variables from the model with no effect on the model’s practicality. For now, Model 4 is the best model we have produced. We are still searching for a model that provides a better value but due to the distribution of the data, this may not be achievable. Our next proposed model will use Model 4 but include second-order terms for Bal\_Total and AGE to test if these variables effects are non-linear and the interactions between these two variables, this will be Model 5:

**Model 5:**

**Hypothesis:**

**Output:**

**A screenshot of a computer

Description automatically generated** **A screenshot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A graph of a graph showing a number of black dots

Description automatically generated**

**Partial F-Test:**

|  |  |
| --- | --- |
| Sum of Squares | 660.29430741 |
| Numerator DF | 3 |
| F Ratio | 60.017463224 |
| Prob > F | 2.534678e-38 |
|  |  |

From Model 5’s output, we have a P-value of < 0.0001 indicating we would reject the null hypothesis, the inclusion of to the model is statistically significant. Our test statistic (F Ratio) value of 164.939 while lower than the previous models is still high enough to indicate model viability. With this model, we have also produced our highest value yet at 0.168. From the parameter estimates we see that the addition of these second-order terms all have low individual P-values and are all significant to the performance of the model; however the additions of these variables has increased the P-value for the AGE variable. On the Actual vs Predicted plot there is continued improvement to the clustering of the data being tighter and closer fitted to the line than previous models. Finally, running a partial F-Test also produces a low p-value, again indicating the significance of these variables’ inclusion. We will test one final model which will build off of Model 5 and have the interactions of the Balance Total and Customer Age on the Offer variable from Model 5. This will be Model 6, shown below:

**Model 6:**

**Hypothesis:**

**Output:**

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

A graph with black dots and red lines

Description automatically generated

A screenshot of a computer

Description automatically generated

**Partial F-Test:**

|  |  |
| --- | --- |
| Sum of Squares | 386.90407187 |
| Numerator DF | 5 |
|  |  |
| F Ratio | 21.395764965 |
| Prob > F | 2.572568e-21 |

For our final model, we have again gained some improvements with the interactions on the Offer variable included. With a P-value of < 0.0001 we would reject the null hypothesis, meaning that the inclusion of these variables is statistically significant. Across all models we haven’t found a great value, likely due to the large clustering of data around the origin with very little variability on the x axis and a large degree of variability on the y-axis. But, this model is technically viable. We have increased our value again with this model to 0.179. and we still have a high test statistic value of 115.157. Conducting a partial F-test provides a P-value of < 2.5725e-21 so again we have shown that these variables are statistically significant. With the parameter estimates table we propose our final model to be:

**Conclusion:**

With the distribution of the data collected, it may prove difficult to build a model that accounts for a high degree of variance. Much of the data is clustered near the origin with the median revenue generated per account being only $1.17 over a six-month period. Nevertheless, the final model proposed is viable despite much of the variance being unaccounted for. We were able to narrow the scope of our model to the most crucial variables necessary for making revenue predictions, allowing for a more simplified model to be used rather than an overly complex one. There is a possibility for further improvements that we would propose. It is possible that interactions with a different variable or interactions with all variables may produce a more effective model though this was the best model we could produce while still meeting the deadline. This model does provide some additional insights that may benefit the bank. We have found the accounts that have the largest positive effect on revenue are Credit Card accounts, Personal Loan accounts, and Insurance accounts. If the bank were looking to grow revenue, promoting these accounts may help grow the bank.