Case Analysis on Nils Baker

Chen Ziao U1420681G

Critical Issue

After talking with a woman who were frustrated on difficulties in money exchange during travelling, Nils Baker, vice president of a regional retail bank in U.S., has identified a possibility of potential demand in more physical bank branches to facilitate bank's checking account service. To test the idea, he asked Anna Gruer to gather some data and perform some analysis to support it.

Problem Statement

Would the presence of a physical bank branch in a Metropolitan Statistical Area increase the demand of checking accounts?

Analysis

The data of this case was on 120 Metropolitan Statistical Areas. It was in excel format and each record comprised of 4 features. Table 1 shows the detailed description of these features. All analysis was performed using *Python Jupyter Notebook* with packages *numpy*, *pandas* and *statsmodel*. The analysis was divided into 4 sections: Data Exploration, Data Preprocessing, Correlation Analysis, Linear Regression.

Feature	Type	Description
ID	Numeric	ID of the record
Total Household in Area	Numeric	Total number of households in a MSA
Households with Account	Numeric	Total number of households that have a checking account in Nils Baker's Bank in a MSA
Inside/Outside Footprint	Categorical	Indicate whether there is a physical branch of Nil Baker's bank in a MSA

Table 1. Feature Descriptions

Data Exploration

Data exploration was first performed on the raw data to gain some basic insights and understanding of MSRs data. Figure 1 showed a summary of its numeric features.

	ID	Total Households in Area	Households with Account
count	120.000000	1.200000e+02	120.000000
mean	60.500000	1.626027e+05	1992.258333
std	34.785054	2.769554e+05	3301.492956
min	1.000000	1.799000e+03	13.000000
25%	30.750000	2.886075e+04	326.000000
50%	60.500000	4.608250e+04	565.000000
75%	90.250000	1.524350e+05	2092.000000
max	120.000000	1.772960e+06	18133.000000

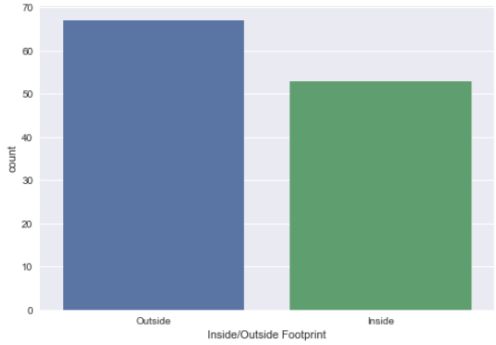

Figure 1. Summary of MSAs Data Set Statistics

From Figure 1, it could be observed that feature "ID" was just a sequence from 1 to 120, which failed to provide us with any useful insights and it should be removed. For feature "Total Households in Area" and "Households with Account", they had large standard deviation relative to its mean and the max values of them were very far away from their 75th percentiles. Therefore, there might be some outliers and this issue would be addressed in data preprocessing section. From the count plot on the right, it could be concluded that the number of "Outside" and "Inside" did not differ much, which indicated the data was relatively balanced for later regression analysis and no oversampling or undersampling was required. There were 67 MSAs (55.8%) without a branch in and 53 MSAs (44.2%)

with a branch. This shows that Nils Baker's bank might need to roughly double its current number of branches to cover all 120 MSAs.

Data Preprocessing

As the previous section suggested, feature "ID" was removed from the dataset. There were two preprocessed data sets generated at this stage. One was for Correlation Analysis and another one was for Linear Regression Analysis.

For Correlation Analysis, since features "Total Households in Area" and "Households with Accounts" were not intuitive enough to obtain insights, they were combined to generate a new numeric feature called "Coverage Ratio" by using the formula (Coverage Ratio = Households with Account / Total Households in Area). For convenience, the feature "Inside/Outside Footprint" was renamed as "Branch". After that, all features were removed except for "Coverage Ratio" and "Branch". Label encoding was applied on feature "Branch" to convert it into 1 and 0, where 1 meant there was a branch and vice versa. Figure 2 below showed the first 5 records of the preprocessed data set.

	Branch	Coverage Ratio
0	0	0.009906
1	0	0.010814
2	0	0.011294
3	1	0.019534
4	0	0.005918

Figure 2. Preprocessed MSAs Data Set (for Correlation Analysis)

For Linear Regression, feature "Branch", "Total Households in Area" and "Households with Accounts" were kept. For feature "Branch", preprocessing was already done above. For the remaining features, min-max normalization was performed on other two features to ensure all values ranging from 0 to 1. Figure 3 showed the first 5 records of the preprocessed data set.

	Total Households in Area	Households with Account	Branch
0	1.000000	0.968543	0
1	0.758491	0.802097	0
2	0.541247	0.597903	0
3	0.523089	1.000000	1
4	0.503735	0.291280	0

Figure 3. Preprocessed MSAs Data Set (for Linear Regression)

Correlation Analysis

To detect whether there was a relationship between the physical presence of a branch and checking accounts coverage ratio, a Pearson correlation coefficient analysis was performed. Figure 4 showed the result of coefficient analysis and Figure 5 showed the scatter plot of preprocessed data.

	Branch	Coverage Ratio
Branch	1.000000	0.152333
Coverage Ratio	0.152333	1.000000

Figure 4. Pearson Correlation

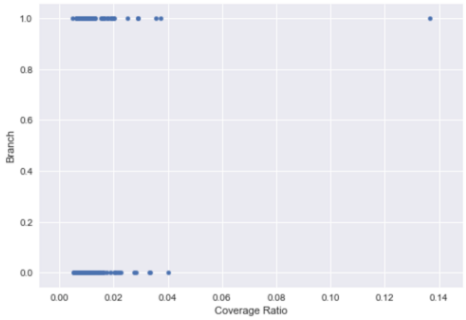

Figure 5. Scatter Plot

From Figure 4, it showed that features “Coverage Ratio” and “Branch” had a positive correlation, but the correlation was only around 0.15, which was not very strong. However, according to Figure 5, it seemed that most “Coverage Ratio” ranged from 0 to 0.04 and in this range, the data was quite evenly distributed between “Branch” value 1 and 0 except for one outlier at point (0.1366, 1). Therefore, this outlier might be the main cause for the positive correlation. To remove the outlier effect, another correlation analysis was performed after removing point (0.1366, 1).

	Branch	Coverage Ratio
Branch	1.000000	0.120973
Coverage Ratio	0.120973	1.000000

Figure 6. Pearson Correlation (Outlier Removed)

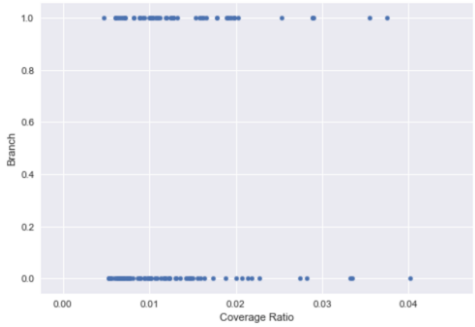

Figure 7. Scatter Plot (Outlier Removed)

From Figure 6, it showed that the positive correlation between two features decreased, but the decrement amount was not very significant. From Figure 7, the scatter plot showed that when there was a branch, “Coverage Ratio” was clustered at low value range 0.05 to 0.15 and when there was no branch, values did not cluster into any range. Hence, there was a potential positive relationship between two features.

Linear Regression Analysis

Another way to identify relationship between variables was using linear regression analysis. For this case, feature “Total Households in Area” and “Branch” were independent variables while feature “Households with Account” was dependent variable (target variable). Figure 8 shows the result of Regression.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0008	0.011	0.074	0.941	-0.021	0.023
Total Households in Area	1.0835	0.046	23.535	0.000	0.992	1.175
Branch	0.0227	0.014	1.576	0.118	-0.006	0.051

R-squared:	0.834
Adj. R-squared:	0.831

```
const                       9.414671e-01
Total Households in Area    3.441288e-46
Branch                      1.177931e-01
```

Figure 8. Linear Regression Result (P Value is the Rightmost Snippet)

From Figure 8 result, the coefficient for both independent variables were both positive. However, the coefficient for “Total Households in Area” was much higher than “Branch” and p value of it was also much lower. It was reasonable because normally the more households in an area, the more checking accounts would be opened. The p-value for “Branch” is greater than 0.05, so the coefficient was not statistically significant. Based on R-squared and Adjusted R-squared, they were both large and close to one, which might be due to the variable “Total Households in Area”. In short, the variable “Branch” had a very weak positive relationship with variable “Households with Account”.

Conclusion

From the Correlation Analysis and Linear Regression Analysis, it could be concluded that the existence of physical branch in a MSA would increase the number of households which opened a checking account in Nils Baker’s bank. However, the impact was relatively weak and just creating a physical branch would not significantly improve checking account business. There might be many

other factors that determined the number of household accounts in a MSA and physical branch would be just one of them.

Recommendations

Since there was no very significant positive relationship between existence of physical branch and number of checking accounts (or checking accounts coverage), the decision on whether setting up physical branches at different regions should be made after further analysis by considering more factors such as cost to set up a branch, economic level (GDP) and exact location of the branch in a MSA.