**Case Analysis on Nils Baker**

# Critical Issue

After talking with a woman who were frustrated on difficulties in money exchange during travelling, Nils Baker, who is a 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 this assumption, 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 for the following case analysis is from 120 Metropolitan Statistical Areas. It is in excel format and each record comprises of 4 attributes. All analysis is performed using Python Jupyter Notebook with packages numpy, pandas and scikit-learn. Figure 1 shows the first 5 records of the data set. The analysis is divided into 5 sections: Data Exploration, Data Preprocessing, Correlation Analysis, Simple Linear Regression and Logistic Regression.



Figure 1. First Five Records of MSAs Data Set

|  |  |  |
| --- | --- | --- |
| 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

Before any further processing of data, simple data exploration is performed on the raw data to gain some basic insights and understanding of MSRs data. Figure 2 shows a summary of data set numeric features.



Figure 2. Summary of MSAs Data Set Statistics

From Figure 1 and Figure 2, it can be observed that feature “ID” is just a sequence from 1 to 120. Therefore, it cannot provide us with any further insights and it should be removed. For feature “Total Households in Area” and “Households with Account”, they have quite large standard deviation relative to its mean and the max of them is very far away from their 75th percentiles. Therefore, there may be some outliers in the data and this issue will be addressed in data preprocessing section. A count plot for the last feature “Inside/Outside Footprint” is shown below in Figure 3.



Figure 3. Count Plot for Feature “Inside/Outside Footprint”

From Figure 3, it can be concluded that in the data set, the number of “Outside” and “Inside” does not

## Data Preprocessing

## Correlation Analysiss

## Logistic Regression

## Decision Tree

# Recommendations

# Conclusion