United States Bill Payment Machine Location Modeling

Jiawen Liu[1]
[1]Management Information System, Oklahoma State University, Stillwater, OK 74078

Introduction

Paying bills without a bank account can be challenging. The company sponsoring this project is a provider of kiosks which facilitates customers to pay their bills via cash, checks, and credit cards through an automatic process. The bills that can be paid include utilities, phone and insurance.

This project is aimed to build a predictive model to estimate the total number of payment transactions in a retail stores on an annual basis. JMP® Pro 10 is used to build predictive models; more precisely data preparation uses JMP® Scripting Language.

The dataset provided by the sponsoring company contains information about kiosks located in USA. Final dataset used in this poster contains information about 78 kiosks only. In this dataset, market related variables such as demographics, direct bills, and share influence variables, such as store type, brand, and competitors all collected within 1 mile, 3 miles, and 5 miles radius of each kiosk. The data is then standardized and transformed for later analysis.

Regression and Factor Analysis followed by factor-bases scales are used in this project. As per our NDA with the sponsoring company, all transaction related numbers have been masked.

Methods

Step I Data Preparation

Variables are measured in different scales. Several demographic variables, like “populations”, have wide range of variation and large values, which can have major effects on modeling. Also, they are highly skewed. These numeric variables need to be transformed and standardized to improve constancy and normality. Log transformation is used for right skewed variables; square root transformation is used for left skewed variables. Next, column standardization is used for shaping the variables in a evenly range. Nominal variables are recoded into dummy variables so that they can be used as numeric inputs.

[1] is an example of recoding categorical variables into binary variables.

[1] new column(“New_Name”, Numeric, formula(if(is missing(Category_Var()),0,1)));

[2] is an example of normal standardizing the demographic variables.

[2] new column(“New_Name2”, Numeric, formula(col standardize(Demographic_Var)));

[3] is an example of Log transformation.

[3] new column(“New_Name3”, Numeric, formula(log(Variable+1)));

[4] is an example of Square Root transformation.

[4] new column(“New_Name4”, Numeric, formula(sqrt(Variable)));

Step II Modeling

Regression Modeling

Linear regression model is used by considering all transformed and standardized variables as inputs, stepwise method as selection method, and 90% significant confidence level as start and stop probability. In the results, no variable shows potential multicollinearity problems. In the final model, 10 significant variables are retained with the overall adjusted R-Square of 0.78.

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.073112	0.049159	123.54	<.0001*
Multiple	1.5167591	0.345043	4.40	<.0001*
Hispanic1	0.1725023	0.06501	2.65	0.0099*
African_American1	-0.1213	0.063096	-1.92	0.0587
Competitors1	0.1912658	0.073721	2.59	0.0116*
Caucasian3	-0.502386	0.076468	-6.57	<.0001*
Divorced3	0.8342341	0.139447	5.98	<.0001*
Competitors3	-0.299831	0.082243	-3.65	0.0005*
Divorced5	-0.644257	0.139876	-4.61	<.0001*
No_of_Utility	0.2687921	0.052717	5.10	<.0001*

Figure 1.1 Regression Parameter Estimate

As seen from figure 1.1, 6 variables have positive effects, and 4 variables have negative effects. “Multiple” has the most positive impact. It represents that if currently there is one additional kiosk at the specific location, the total number of transaction will increase by 1.52 (Target has been masked, thought it’s a small number, in fact, it is a big difference), while holding other variables constant.

Three-Factor Regression Modeling

1. MSA test is used to eliminate the unsuitable variables for factor analysis. All variables with a MSA of less than 0.5 are rejected. Overall MSA for all retained variables is 0.72.

2. In factor analysis, the number of components is decided based on cumulative variance explained by all variables and the eigenvalue. Varimax rotation method is used for factor rotations. While determining the number of factors, variables which have high cross-loadings on multiple factors are removed. Three factors (3F) are retained in the final analysis (Table 1.1) and these 3 factors account for 74 % of cumulative variance in all of the variables.

3. Maximum Likelihood significant test shows three factors are sufficient (Figure 1.2).

4. An average three factors scale Regression model is built with adjusted R-Square of 0.51. In the parameter estimates, all three scales show significant result and no potential multicollinearity problems (Figure 1.3). Scale 3 shows negative impact, and other two scales show positive impact. Scale 2 has the most effects on total transactions.

Factor	Factor1	Factor2	Factor3
Variables	Divorced3, Divorced5	Multiple, Competitors1, No_of_Utility	African_American 1, Caucasian3

Table 1.1 Three Factors

Test	DF	ChiSquare	Prob>ChiSq
H0: 3 factors are sufficient.	3.000	3.387	0.3358
HA: more factors are needed.			

Figure 1.2 Maximum Likelihood Significant Test

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.1348862	0.06955	88.21	<.0001*
Scale1	0.2436159	0.07123	3.42	0.0010*
Scale2	0.7652918	0.138396	5.53	<.0001*
Scale3	-0.633083	0.160099	-3.95	0.0002*

Figure 1.3 Three-factor regression parameter estimate

Results

Model	R-Square	Adjusted R-Square	RMSE	Significant test
Regression	0.81	0.78	0.41	Significant
3F Regression	0.53	0.51	0.61	Significant

Table 1.2 Model Comparison

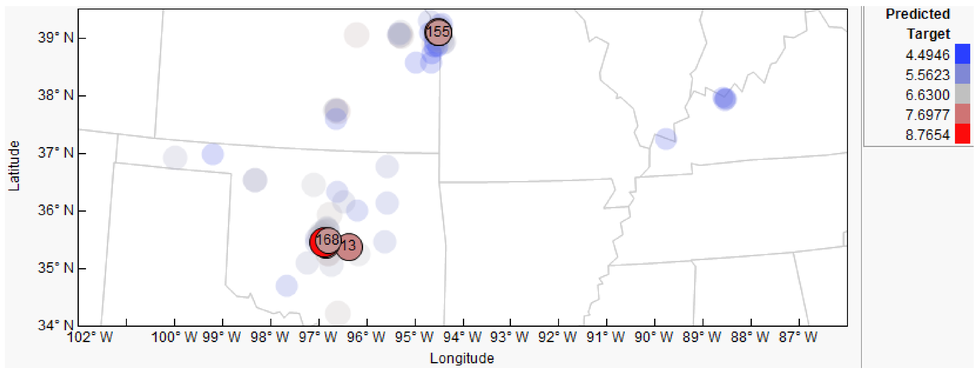

Figure 1.4 Prediction of high volume location

Discussion

•Seen from table 1.2, the Regression model with the transformed performs better (higher adjusted R-Square and lower RMSE) than the factor-based scales.

•Figure 1.4 is generated by JMP® Pro 10 Bubble Plot. Marked “ID” kiosks have been employed for more than 24 moths and predicted as having high volume location.

•The regression model shows that having multiple kiosks in the retail stores could increase the average number of total transactions. Higher percentage of Hispanics in the geographical neighborhood tend to increase the number of transactions, while higher percentage of Caucasians tend to reduce the number of transactions.

Reference

1. SAS Institute Inc. 2012. *JMP® 10 Scripting Guide*. Cary, NC: SAS Institute Inc.

Acknowledgement

I thank Dr. Goutam Chakraborty for his guidance on this project. I also thank the sponsoring company that provided the data and who wishes to remain anonymous.