Hypothesis Testing

Enqun Wang (EW), Yiyan Zhou (YZ)

April 26, 2016

Exploratory Data Analysis

According to the exploratory data analysis, we decide to first elect variables as follows,

Product Factor	Promotion Factor	Platform Factor	Market Factor
1. Rate of Return	1. LB Received	1. Balance	1. R.007
2. Term	2. LB Used	2. Capital Inflow	
3. TZD Account			

We use the centered data to test the significance of the variables, and find the significant ones.

1. Increase.Rate.5Day ~ Rate.of.Return

```
##
## Call:
## lm(formula = df$trans.Increase.Rate.5Day ~ df$trans.Rate.of.Return)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8454 -0.7421 -0.1314  0.6312  3.4207
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               8.331e-19  1.155e-01   0.000    1.000
## df$trans.Rate.of.Return -5.334e-03  1.162e-01  -0.046    0.964
##
## Residual standard error: 1.007 on 74 degrees of freedom
## Multiple R-squared:  2.845e-05,  Adjusted R-squared:  -0.01348
## F-statistic: 0.002106 on 1 and 74 DF,  p-value: 0.9635
```

The p value of the test is 0.9635, which is greater than 0.05. We should not reject the null hypothesis and conclude that the coefficient of Rate.of.Return is not significant.

2. Increase.Rate.5Day ~ Term

```
##
## Call:
## lm(formula = df$trans.Increase.Rate.5Day ~ df$trans.Term)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7531 -0.7831 -0.1170  0.6342  3.1287
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    2.550e-17  1.141e-01   0.000    1.000
## df$trans.Term -1.564e-01  1.148e-01  -1.362    0.177
##
## Residual standard error: 0.9943 on 74 degrees of freedom
## Multiple R-squared:  0.02445,    Adjusted R-squared:  0.01127
## F-statistic: 1.855 on 1 and 74 DF,  p-value: 0.1773
```

The p value of the test is 0.1773, which is greater than 0.05. We should not reject the null hypothesis and conclude that the coefficient of Term is not significant.

3. Increase.Rate.5Day ~ TZD.Account

```
##
## Call:
## lm(formula = df$trans.Increase.Rate.5Day ~ df$trans.TZD.Account)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7776 -0.7596 -0.0298  0.5272  3.3339
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.943e-17  1.102e-01   0.000  1.00000
## df$trans.TZD.Account -2.976e-01  1.110e-01  -2.682  0.00903 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9611 on 74 degrees of freedom
## Multiple R-squared:  0.08857,    Adjusted R-squared:  0.07625
## F-statistic: 7.191 on 1 and 74 DF,  p-value: 0.009031
```

The p value of the test is 0.009031, which is less than 0.05. We should reject the null hypothesis and conclude that the coefficient of TZD.Account is significant.

4. Increase.Rate.5Day ~ LB.Received

```
##
## Call:
## lm(formula = df$trans.Increase.Rate.5Day ~ df$trans.LB.Received)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7086 -0.8061 -0.1410  0.6606  3.2724
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.240e-17  1.129e-01   0.000    1.000
## df$trans.LB.Received -2.098e-01  1.137e-01  -1.846    0.069 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9843 on 74 degrees of freedom
## Multiple R-squared:  0.044,  Adjusted R-squared:  0.03108
## F-statistic: 3.406 on 1 and 74 DF,  p-value: 0.06896
```

The p value of the test is 0.04645, which is less than 0.05. We should reject the null hypothesis and conclude that the coefficient of LB.Received is significant.

5. Increase.Rate.5Day ~ LB.Used

```
##
## Call:
## lm(formula = df$trans.Increase.Rate.5Day ~ df$trans.LB.Used)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7526 -0.7464 -0.1423  0.6034  3.4414
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.926e-17  1.148e-01   0.000    1.000
## df$trans.LB.Used  1.066e-01  1.156e-01   0.922    0.359
##
## Residual standard error: 1.001 on 74 degrees of freedom
## Multiple R-squared:  0.01136,    Adjusted R-squared:  -0.001997
## F-statistic: 0.8505 on 1 and 74 DF,  p-value: 0.3594
```

The p value of the test is 0.1695, which is greater than 0.05. We should not reject the null hypothesis and conclude that the coefficient of LB.Used is not significant.

6. Increase.Rate.5Day ~ Balance

```
##
## Call:
## lm(formula = df$trans.Increase.Rate.5Day ~ df$trans.Balance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8752 -0.7416 -0.1388  0.6520  3.3705
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.082e-17  1.154e-01   0.000    1.000
## df$trans.Balance  3.499e-02  1.162e-01   0.301    0.764
##
## Residual standard error: 1.006 on 74 degrees of freedom
## Multiple R-squared:  0.001224,   Adjusted R-squared:  -0.01227
## F-statistic: 0.0907 on 1 and 74 DF,  p-value: 0.7641
```

The p value of the test is 0.8785, which is greater than 0.05. We should not reject the null hypothesis and conclude that the coefficient of Balanec is not significant.

7. Increase.Rate.5Day ~ Capital.Inflow

```
##
## Call:
## lm(formula = df$trans.Increase.Rate.5Day ~ df$trans.Capital.Inflow)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6154 -0.7388 -0.0793  0.6976  2.1425
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.907e-17  1.034e-01   0.000        1
## df$trans.Capital.Inflow  4.457e-01  1.041e-01   4.283 5.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9012 on 74 degrees of freedom
## Multiple R-squared:  0.1986, Adjusted R-squared:  0.1878
## F-statistic: 18.34 on 1 and 74 DF,  p-value: 5.467e-05
```

The p value of the test is 5.467e-05, which is less than 0.05. We should reject the null hypothesis and conclude that the coefficient of Capital.Inflow is significant.

8. Increase.Rate.5Day ~ R.007

```
##
## Call:
## lm(formula = df$trans.Increase.Rate.5Day ~ df$trans.R.007)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62852 -0.78853 -0.06729  0.70667  3.14317
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.368e-18  1.113e-01   0.000   1.0000
## df$trans.R.007 -2.678e-01  1.120e-01  -2.391   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.97 on 74 degrees of freedom
## Multiple R-squared:  0.07169,    Adjusted R-squared:  0.05915
## F-statistic: 5.715 on 1 and 74 DF,  p-value: 0.01936
```

The p value of the test is 0.03061, which is less than 0.05. We should reject the null hypothesis and conclude that the coefficient of R.007 is significant.

To sum up, LB.Received, Capital.Inflow, TZD.Account and R.007 are four significant variables.

Considering the above four significant the variables, we do model selection by forward stepwise.

```
## Subset selection object
## Call: regsubsets.formula(df$trans.Increase.Rate.5Day ~ df$trans.LB.Received +
##     df$trans.Capital.Inflow + df$trans.TZD.Account + df$trans.R.007,
##     data = df, method = "forward")
## 4 Variables  (and intercept)
##                           Forced in Forced out
## df$trans.LB.Received        FALSE      FALSE
```

```
## df$trans.Capital.Inflow     FALSE      FALSE
## df$trans.TZD.Account        FALSE      FALSE
## df$trans.R.007              FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: forward
##          df$trans.LB.Received df$trans.Capital.Inflow df$trans.TZD.Account
## 1  ( 1 ) " "                  "*"                     " "
## 2  ( 1 ) " "                  "*"                     "*"
## 3  ( 1 ) "*"                  "*"                     "*"
## 4  ( 1 ) "*"                  "*"                     "*"
##          df$trans.R.007
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) "*"
```

It seems that the capital inflow is the most significent. We can now select the best model based on the voice of customer.

The importance of the four variables are as followed:

Importance	Variable
1	Capital Inflow
2	Lucky Bag Received
3	R 007
4	The Financial Product TZD Account