**Analysis of key factors that influence credit card activity using Regression**

Arpita Shrivas, Yash Patel, Sai Teja, Rakshitha Nirmala

**Step 1: Select the independent variables that you believe may be related to the dependent variable**

**Considered below factors for identifying variables:**

* Included less number variables to decrease the probability Type1 errors
* Objective is to determine whether the factors are playing a significant role in the model or not
* Tried to include the variables which in theory affect the dependent variable

**Independent Variables:**

* Income
* Credit Debt
* Other Debt
* Age – RATIO LEVEL VARIABLE
* Card Items Monthly
* Card Tenure

**Observations:**

* Another variable ‘Debt to income ratio’ could have also played an important role in the model. However, we observed that R square did not change significantly for the model.
* It also affected the p value of household income in an adverse way (Fig 1). The p value for household income dropped by 0.886. Income being an important factor in credit card, we rejected the ‘Debt to income ratio’ variable from the model.
* We can notice here that, ‘Debt to income ratio’ does not look like an independent variable. It is correlated to other variables - Credit Debt, Other Debt and income. The goal, therefore, is to have minimal or lesser multicollinearity.
* Another variable ‘Card Tenure’ did not affect the R square value much. However, the p- value for income became significantly ‘high’ (fig 3) when removed from the model. Thus, we included the ‘Card tenure' in the model.
* Hence, the final prediction model can be represented as:

**YExpense = B0 +B1Xincome + B2XCredit Debt+ B3Xother Debt + B4XAge + B5Xitems + B6XTenure**

**Step 2: Compute all the coefficients and other statistics**

A picture containing text, receipt

Description automatically generatedFig 1 - Includes Debt to income ratio and Card tenure – **REJECTED MODEL**

Table

Description automatically generated

Fig 2 – removed Debt to income ratio – **FINAL MODEL**

**Graphical user interface, application, table

Description automatically generated**

Fig 3 – rejected card tenure variable which led to high p value of income (7.98). Hence, considered Fig 2 as final predictive model.

**Step 3: Evaluate the Model**

**Standard error of estimate**

As we can see, the standard error is 1949.601. It appears that the standard error of estimate is quite large. However, because there is no predefined upper limit on Standard limit, it is often difficult to assess the model this way. In general, the standard error of estimate cannot be used as an absolute measure of the model’s utility. Nonetheless, it is useful in comparing models. If we would have had models from which to choose, the one with the smallest value of standard error should generally be the one used.

**Coefficient of Determination**

* R-square gives the percentage variation in y explained by x-variables. The usefulness of R2 is its ability to find the likelihood of future events falling within the predicted outcomes. Our model has an approx. R2 of 36.7% which means 36% of the variation in independent variables contributes to credit card balance.
* Hence, the selected variables contribute 36.7% to the credit card balance. Adding more variables can increase R2.

**Step 4: Check significance**

* To test the validity of the regression model, we specify the following hypotheses:
* H0: β1 =β2 =···=βk =0

H1: At least one βi is not equal to 0.

* F=Mean Regression Sum of Squares (MRS)/Mean Squared Error (MSE)
* F= 18351/ 38009=0.482
* If the p-value is small (less than [alpha level](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-an-alpha-level/)), we can reject the null hypothesis.

**Conclusion:**

* We can see from the p values below that they all are greater than alpha value 0.005. Hence, we do not reject the null hypothesis. The sample averages are significantly different from each other. The observed differences among the sample averages could not reasonably be due to random chance alone. The result is statistically significant.

|  |  |
| --- | --- |
| **Variable** | **P-value** |
| HHIncome | 2.21071 |
| CreditDebt | 0.06801884 |
| OtherDebt | 0.34814114 |
| Age | 0.10528543 |
| CardItemsMonthly | 0 |
| CardTenure | 0.53797708 |

* As explained earlier, it can also be said that another variable ‘Debt to income ratio’ could have also played an important role in the model. However, we observed that R square did not change significantly for the model.
* It also affected the p value of household income in an adverse way (Fig 1). The p value for household income dropped by 0.886. Income being an important factor in credit card, we rejected the ‘Debt to income ratio’ variable from the model.
* We can notice here that, ‘Debt to income ratio’ does not look like an independent variable. It is correlated to other variables - Credit Debt, Other Debt and income. The goal, therefore, is to have minimal or lesser multicollinearity.
* Another variable ‘Card Tenure’ did not affect the R square value much. However, the p- value for income became significantly ‘high’ (fig 3) when removed from the model. Thus, we also included the ‘Card tenure' in the model.
* We can say that Household Income, Credit Debt, Other Debt, Age, Card items monthly and Card tenure contribute to 36.7% of the credit card balances. Hence, they can be considered as key factors affecting credit card activity.
* We can also conclude that one of the variable ‘AGE’ which is a ratio level variable has a negative coefficient which means as the age increases, the credit card balance tends to decrease. The closer the coefficient is to -1.0, the stronger the negative relationship will be which means they are significantly inversely proportional as the value is -3. 71. We can see that people tend to use less credit cards as they age.
* As we can see, Ratio level variable – Age is in relation with one of the categorical variable in the dataset ‘Retired’. Hence, including both variables could have created some level of irrelevancy in the model. Similarly, age and ‘marital status’ can be related to each other hence only one variable was considered in the model.