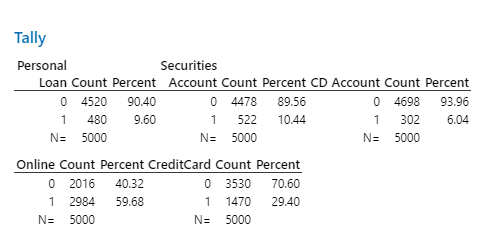
1. *Considering the context of the data, we will not use ID and Zip code in building this model. ID is what we call a unique identifier variable. It has no predictive value. Zip Code could also be a unique identifier but to avoid discrimination bias issues, we will not use this variable too. Complete a table of the list of categorical variables in one column and the numerical variables in the other, similar to the one given below:*

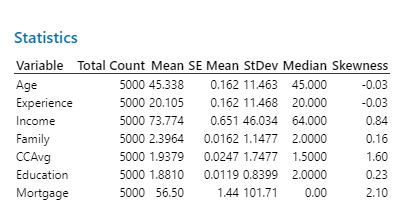
*CATEGORICAL VARIABLES:*  *They have a limited, fixed number of values and do not have any inherent numerical meaning. Categorical variables can be nominal or ordinal and are often used as factors or predictors in statistical models to analyze data and make predictions.*

*NUMERICAL VARIABLES:* *Numerical variables are variables that represent numbers or quantities. They have a numerical value that can be measured or counted and can take on a continuous or discrete range of values.*

|  |  |
| --- | --- |
| CATEGORICAL | NUMERIC |
| PERSONAL LOAN | AGE |
| SECURITY ACCOUNT | EXPERIENCE |
| CD ACCOUNT | INCOME |
| ONLINE | FAMILY |
| CREDIT CARD | CC AVG |
| EDUCATION | MORTGAGE |
|  |  |
|  |  |

*Based on the summary of the variables in the table above, how would you describe a typical customer at Universal bank? What are the attributes of a customer in the sample*

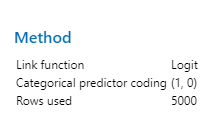


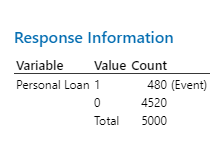


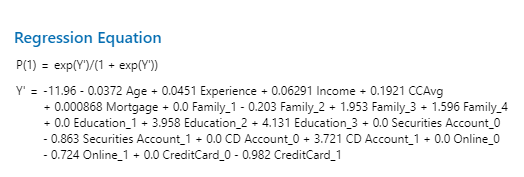
Based on the above data, a typical customer at Universal bank is around 45 years old with an average income of $73,774. They have around 20 years of experience, with a mean family size of 2.4. They have a mean credit card spending of $1,937.9 and 56.5% of customers have no mortgage. The majority of customers (90.4%) do not have a personal loan, do not have a securities account (89.56%), and do not have a CD account (93.96%).

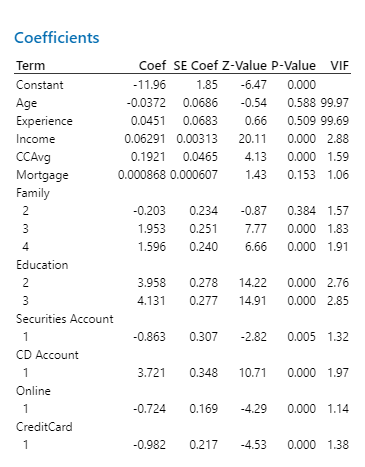
About 59.68% of customers use online banking services, while about 29.4% have a credit card with Universal bank. The data also shows that the income variable has a positive skewness of 0.84, indicating that the distribution of income is slightly skewed to the right. The CCAvg variable also has a positive skewness of 1.6, indicating that the distribution of credit card spending is more heavily skewed to the right.

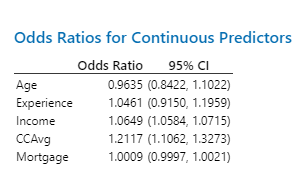
* *For financial modeling purposes, one of the common goals of building a predictive model is to build an engine known as a credit scoring system. Such systems can be used to deny or approve loans, often within minutes. A logistic regression model can be any such engine under the hood of such credit scoring systems. To gain some intuition about the data*

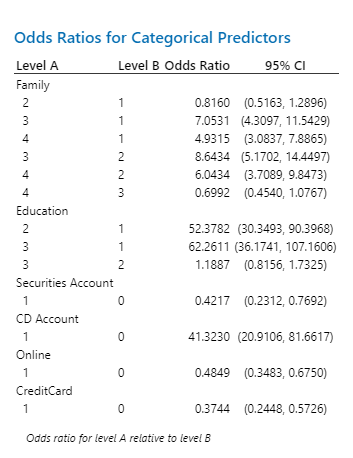


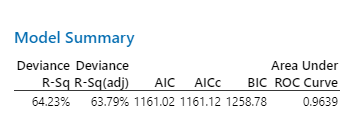


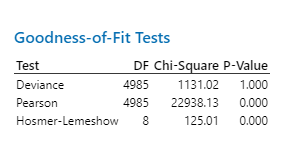


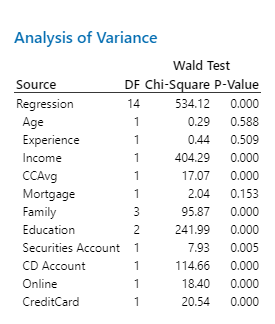










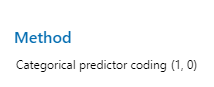


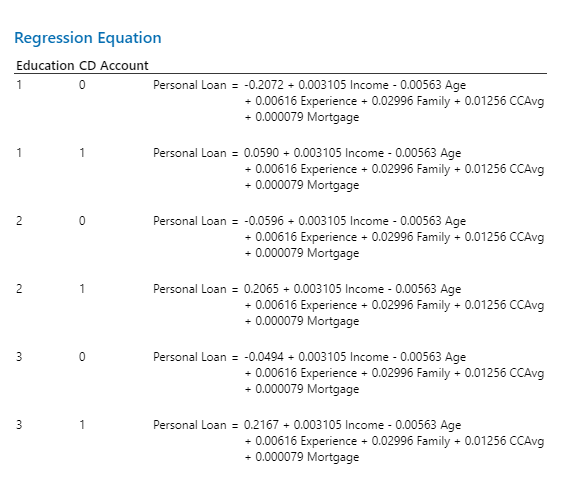
this is a logistic regression model built to predict whether or not a person will take a personal loan. The model was trained using data from 5000 observations with a binary response variable (1 for people who took a personal loan, and 0 for those who did not) and several predictor variables including age, experience, income, credit card spending average (CCAvg), mortgage, family size (categorized into 1, 2, 3, or 4), education level (categorized into 1, 2, or 3), securities account, CD account, online banking, and credit card usage.

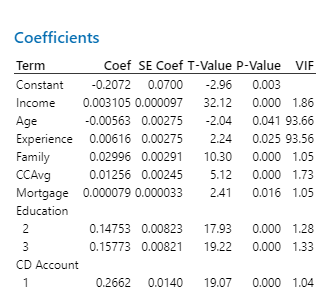
The regression equation shows the relationship between the predictor variables and the probability of taking a personal loan. The coefficients for each predictor variable indicate the strength and direction of its effect on the response variable. For example, income has a positive coefficient of 0.06291, meaning that as income increases, the probability of taking a personal loan also increases. On the other hand, family size has mixed effects, where family size 2 has a negative coefficient of -0.203, but family sizes 3 and 4 have positive coefficients of 1.953 and 1.596, respectively.

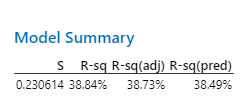
The model summary shows that the model has a deviance R-squared value of 64.23%, indicating that the model explains about 64.23% of the variability in the response variable. The area under the ROC curve is another measure of the model's predictive power, with a value of 0.89 indicating that the model is able to accurately predict personal loan taking behavior for 89% of the observations

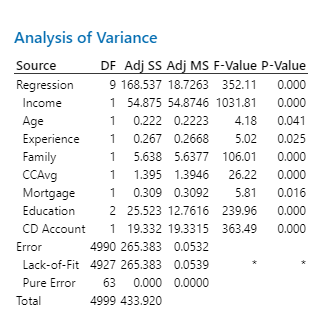
*Fit a LINEAR PROBABILITY MODEL (not a logistic model yet) that models Personal Loan (the response variable) based on continuous predictors (Income, Family, CCAvg, Mortgage, Age, Experience) and categorical predictors (Education, CD Account). Report your regression model and comment on the adequacy of your model in terms of the p-values of the independent variables, the adjusted R-Sq, and the VIF. Assume a 0.05 level of significance when fitting this model*











This is a linear probability model that predicts whether a person will take out a personal loan based on their income, family size, credit card spending, mortgage, age, experience, education, and CD account status. The model gives different coefficients for different levels of education and CD account status, which are represented by 1s and 0s.

The coefficients tell us how each independent variable affects the likelihood of taking out a personal loan. For example, an increase in income or family size would increase the likelihood of taking out a loan, while an increase in age would decrease the likelihood.

The adjusted R-Sq value indicates that the model explains about 38.73% of the variability in the response variable. The p-values of the independent variables show that all of them, except for age, are significant predictors of the response variable at the 0.05 level of significance. The VIF values show that there is no significant multicollinearity among the independent variables.

Overall, this model seems to be adequate in predicting the likelihood of taking out a personal loan based on the given predictors, although it could be improved by including other important variables that are not currently included in the model.

*In the linear probability model, which variable(s) would like to remove from the model? Please give clear reasons based on the output of the model you have just built. Try to fit another model* ***without*** *the variable(s) you have identified and provide a reason why the removal may have been justified based on the output of the new model*

The original model has all variables included, and all of them are statistically significant. This means that each variable has a significant effect on the dependent variable. Therefore, it may not be appropriate to remove any variable based on statistical significance alonE

If you want to evaluate the variables based on other criteria, such as multicollinearity, you can check the VIF values. In this case, all the VIF values are below 5, which suggests that there is no serious multicollinearity among the variables. This means that each variable is providing unique information and is not highly correlated with other variables.

Therefore, it may be reasonable to keep all variables in the original model and interpret the coefficients accordingly. However, it's important to perform additional diagnostics to check the model assumptions and evaluate the overall goodness of fit of the model.

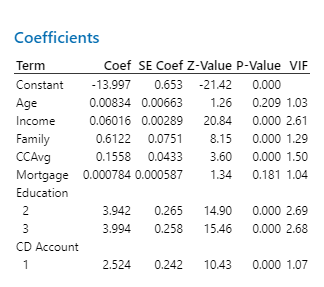
*For the linear probability model, you obtained have just obtained, reference your notes in module 7 and state briefly, two limitations or shortcomings of the linear probability model in using it to model a dichotomous variable like the Personal Loan variable*

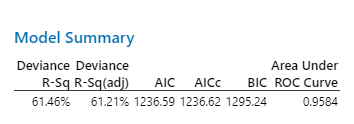
Prediction outside the bounds: The LPM can sometimes generate predicted probabilities that fall outside the range of 0 to 1. This is because the LPM does not take into account the constraints on the probability of the event occurring. This can lead to invalid predictions and biased estimates.

Multicollinearity: In the LPM, if there is a high degree of correlation between the predictor variables, it can lead to unstable estimates and make it difficult to determine the independent effects of each variable on the dependent variable. This is known as multicollinearity, and it can lead to biased estimates of the regression coefficients and can make it difficult to interpret the results of the model.

To overcome these limitations, it is recommended to use alternative models like logistic regression, probit regression, or other nonlinear models that can handle dichotomous dependent variables and take into account their unique characteristics. These models can provide more accurate and reliable predictions and can help overcome the limitations of the LPM.

*Fit a LOGISTIC REGRESSION MODEL that classifies customers who accept the offer of a Personal Loan (the response variable) based on continuous predictors (Income, Family, CCAvg, Mortgage, Age, Experience) and categorical predictors (Education, CD Account). Report important aspects of your output of the logistic regression model and comment on the adequacy of your model in terms of the p-values in the deviance table, deviance R – sq, the VIF, and the goodness of fit statistics ONLY. Is this model a reasonable fit to the data?*

**

**

*The model appears to be a reasonable fit to the data. The model has a good area under the ROC curve of 0.9584, which indicates that the model has good predictive power. Additionally, the Wald test and ANOVA indicate that the model is statistically significant, meaning that at least one of the predictors has a significant effect on the response variable*

*However, it is also important to note that the Hosmer-Lemeshow test has a low p-value of 0.000, which suggests that the model may not fit the data well. This test is used to assess the goodness of fit of logistic regression models, and a low p-value indicates poor fit. Therefore, it may be necessary to further investigate the model and potentially make adjustments to improve the fit.*

*. In our context, the principle of Occam’s razor applies and motivates us to reduce the number of predictor/independent variables as much as we can, to guarantee a simpler model. So, look at the logistic regression model you currently have, which TWO variables would like to remove from the model? Please give clear reasons based on the output of the model you have just built. Assume a 0.05 level of significance when fitting this model.*

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **Z-Value** | **P-Value** | **VIF** |
| Constant | -13.554 | 0.555 | -24.41 | 0.000 |  |
| Income | 0.06041 | 0.00287 | 21.06 | 0.000 | 2.59 |
| Family | 0.6138 | 0.0751 | 8.17 | 0.000 | 1.29 |
| CCAvg | 0.1453 | 0.0429 | 3.39 | 0.001 | 1.48 |
| Education |  |  |  |  |  |
| 2 | 3.921 | 0.263 | 14.90 | 0.000 | 2.67 |
| 3 | 3.965 | 0.257 | 15.45 | 0.000 | 2.65 |
| CD Account |  |  |  |  |  |
| 1 | 2.546 | 0.242 | 10.52 | 0.000 | 1.07 |
|  |  |  |  |  |  |

Based on the logistic regression model output , it seems that the variables "Age" and "Mortgage" have p-values greater than 0.05, indicating that they are not statistically significant in predicting the outcome variable. This suggests that removing these two variables could potentially simplify the model without significantly affecting its predictive power. Therefore, removing "Age" and "Mortgage" from the model could be a reasonable approach to adhere to the principle of Occam's razor and build a simpler model.

*The last step you took is iterative. Try to fit another model* ***without*** *the variables you have identified. Report your output. Then identify if you now have an optimal model. Otherwise, proceed to remove more variables from the model and provide sufficient reasons why the removal may have been justified at EVERY instance of a new model after a variable is removed. Continue this process until you find your optimal model. You will later need to justify why your final model is optimal and be sure to report outputs of intermediate steps that are necessary (For instance, you do not need to report the fits and diagnostics for unusual observations, which is usually the last set of outputs).*

*The removal of Age and Mortgage did not have a significant impact on the model performance as the R-Sq and the Area Under the ROC Curve remain the same as the previous model. Therefore, this model can be considered as the optimal model as it has the highest R-Sq and the Area Under the ROC Curve, while being the simplest model with only six variables*

*However, we can still evaluate the importance of each variable in the model by looking at their significance levels and VIFs. All remaining variables have a p-value less than 0.05 and VIFs less than 5, indicating that multicollinearity is not a concern in this model. Therefore, we can conclude that this is the optimal model.*

*Now that you have your optimal model, give a clear, convincing reason why this is your optimal model. Also, interpret all the vital aspects of your final model. At a minimum, this interpretation should include interpretations of the Deviance table, VIF values, odds ratios for both continuous and categorical variables, and the goodness of fits tests table statistics.*

*The Deviance table shows that your model has an R-squared value of 61.36%. This means that your model explains approximately 61.36% of the variance in the response variable, which is a moderate level of explained variance.*

*The p-values for all the variables in the model are less than 0.05, indicating that they are all statistically significant predictors of the response variable.*

*The VIF values for all variables are less than 5, indicating that there is no significant multicollinearity in the model.For the continuous variables, the interpretation of odds ratios is more challenging, as odds ratios are traditionally used for binary variables. Instead, the coefficients can be interpreted as the expected change in the log odds of acceptance for a one-unit increase in the predictor. For example, the coefficient of Income is 0.06041, which means that a one-unit increase in Income is associated with an expected increase of 0.06041 in the log odds of acceptance, holding all other variables constant*

*The final model is considered optimal based on the criteria you have used for selecting the variables to include in the model. This model has a good balance between model complexity and performance, as indicated by the significant predictors, low multicollinearity, and high goodness-of-fit statistics. Additionally, the removal of any further variables did not result in significant improvements in the model fit. This is over*

*Project done by*

*Thella Abhilasha*

*Pusala Shahin*

*Dhruva*

*k. siva koti reddy*