Q1. The excel data **Happiness** contains variables that are associated with a country’s happiness score. You are expected to build a model to predict a country’s **happiness score**. (total 20 points)

1. Fit a full model to the data and compute the VIF for each predicting variable. Which variable has the highest VIF? (5 points)
2. Remove the variable with the highest VIF from the list of predicting variables and fit a full model with the remaining variables. Which variable has the highest VIF? (5 points)
3. Repeat the process of (2) until all the variables have VIF less than 10. List the names of the remaining variables which VIF are less than 10. (5 points)
4. Use the predicting variables from the final model in (3). Apply the Stepwise model selection method. Present the estimated model coefficients table. (5 points)

Q2. The following questions are based on the `creditdata.sas7dat’. Suppose you would like to develop a predictive model **to predict if a customer will default** on his/her credit card debt. The data contains the following variables:

|  |
| --- |
| Income ($1000): annual income |
| Limit ($): credit limit |
| Rating: credit score |
| Cards: number of credit cards owned |
| Age |
| Education: years of education |
| Gender: Female vs. Male |
| Student: Yes vs. No |
| Married: Yes vs. No |
| Balance ($): account balance not paid off |
| Default: 0: not default; 1: default |

Answer the following questions:

1. **Data Exploration** (total: 25 points)
2. Suppose you would like to explore if there is a relation between the variable **Balance** and **Default**, what will be an appropriate plot to demonstrate their relationship? Justify your answer. (5 points)
3. Create a plot which visualizes the relation between **Balance** and **Default**. (5 points)
4. Suppose you would like to explore if there is a relation between the variable **Cards** and **Default**, what do you think will be an appropriate plot or a statistical method to explore their relationship? Provide the plot or the output of the statistical method. (7.5 points)
5. If you would like to explore if there is a relation between the variable **Married** and **Default**, what do you think will be an appropriate tool, i.e. plot or table, to demonstrate their relationship? Provide the plot or the output of the statistical method. (7.5 points)
6. **Build Predictive Model** (total 40 points)
7. To predict if a customer will default, you are thinking about fitting a linear probability model where the response variable is the probability of the event when Y=1:

**Pr(**

where if not default; if default

is the value of the kth predicting variable for customer i

Do you think this is an appropriate model to predict default? Justify your answer. (5 points)

If you disagree, what transformation of the response variable will you suggest to improve the model? (5 points)

1. Suppose you create a new variable “Education Group”, which classifies the variable into four groups.If you include this new variablein your model and apply the “reference” coding style, how will you **assign values** of the design variables for each Education Group? Fill out the table below: (5 points)

|  |  |  |  |
| --- | --- | --- | --- |
| Education Group | Design Variable 1 | Design Variable 2 | Design Variable 3 |
| 5 ~ 9 Yrs |  |  |  |
| 10 ~ 12 Yrs |  |  |  |
| 13 ~ 16 Yrs |  |  |  |
| Above 16 Yrs |  |  |  |

1. Fit a logistic regression model to predict default using the following predicting variables: (1) Rating (2) Cards (3) Married.

Provide the table “Analysis of Maximum Likelihood Estimates”. (5 points)



Use your fitted model to answer the following questions:

1. If the variable **Rating** increases by one point, what percentage of increase/decrease will you expect for the odds ratio assuming the values of the other predicting variables remain unchanged. (5 points)
2. Write down the fitted logistic equation **if the Marriage status is Yes**. (5 points). Arrange the terms of the equation if necessary.
3. Write down the fitted logistic equation **if the Marriage status is No**. (5 points). Arrange the terms of the equation if necessary.
4. **Based on your answer in b**, if a customer **is married, has only one credit card, and a credit score=600**, what is the **predicted log of the odds**? (5 points).

**Q3. Model Performance Assessment** (total 15 points)

The two columns in the file ‘**CreditCardDefault\_Predict.xlsx’**, **IP\_0** and **IP\_1** represent the predicted probability of **no default** and **default** respectively.

The column “**Default**” represents the actual default status of a credit card holder.

**You will only need to use these three columns (Default, IP\_0, and IP\_1) to answer the following questions.**

Suppose you apply the following the rule to determine if a customer will default or not default:

**If IP\_1 > 0.5, then default. If IP\_1 0.5, then no default.**

1. Fill out the following Classification Matrix in terms of frequency of observations: (5 points)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted Default** | **Predicted No Default** | **Total** |
| **Default** |  |  |  |
| **No Default** |  |  |  |
| Total |  |  |  |

1. Use the Classification Matrix to compute:
2. **True Positive Rate** (sensitivity) (2.5 points)
3. **False Negative Rate** (1 - specificity) (2.5 points)
4. Create a **Lift Chart** and comment on the model prediction performance (5 points)