ISM 6137 - Statistical Data Mining

Assignment 1

Sahil Shah – 19895141

Questions to answer:

1. What variables predict credit scores and by how much?
2. Is there a racial or gender bias on credit score? If so, by how much?

Tasks:

1. Create a predictor table (see sample assignment solutions) with three columns for predictor, expected sign of effect, and a one-sentence rationale for effect. (2 points)



1. Next run relevant models to answer questions 1 and 2 above. No more than 2-3 models in all. Explain the rationale for your models. (1 point)

Text

Description automatically generated

**Model 1 Rationale**: This model includes all the factors that I suspect would have a significant impact on credit score. The variables with an expected effect of “NONE” are not included. Also, number of credit cards is not included as the effect, if any, would be indirect and the research indicates a loose correlation at best.

**Model 2 Rationale:** This model includes the three predictor variables that have significant P-values from the first model, and the ethnicity and gender factor variables. Gender and ethnicity are included here to examine whether there is a measurable, statistically significant credit score bias. I did not include the variables with non-significant p-values for Beta coefficients in model one to better isolate the effect of gender and ethnicity factors. The three variables with significant P-values were included because based on the output of model one, they definitely have an effect on credit score and thus cannot be ignored when examining the effect of gender and ethnicity.

**Model 3** **Rationale**: This model explores the interaction between credit card balance and ethnicity/gender. Bias in credit score can be illuminated by whether two people with the same outstanding balance but of different ethnicity/gender have different credit scores. This was the primary motivation for investigating this model

1. Interpret your models to answer the two questions asked above. (2 points)

**Interpretation:**

Based on the model output I have chosen Model 2 to derive conclusions for the two proposed questions. I chose Model 2 because it illuminates both which variables have the strongest effect on credit score as well as whether there is a gender and/or ethnicity bias in credit score calculation.

The model indicates that the strongest predictor of credit score is whether the person is a student. Being a student correlates to ~98.4 point drop in credit score. Also, it indicates that Income and Balance have a statistically significant effect on credit score, with Beta Coefficient values of ~2.07points/1,000 dollars and ~0.2 points/dollar respectively. Other predictors seem to have either a marginal effect (very small Beta Coefficient) or highly varied (relatively large Std Error) effect, or both, on credit score and thus we can disregard them as relevant predictors for our target variable.

Furthermore, the model indicates that there is no statistically significant effect of gender or ethnicity on credit score. All Beta coefficients values are highly insignificant (smallest P-value is 20X+ the normally accepted threshold value of 0.05) and include zero well within two standard deviations from the mean Beta coefficient. Thus, we can conclude that at least within this data set there is no verifiable bias based on gender or ethnicity on credit score.