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)

**MODEL\_1:**

Call:

lm(formula = Rating ~ Income + Age + Student + Education + Balance +

Married, data = CreditRating)

Residuals:

Min 1Q Median 3Q Max

-80.882 -10.906 4.167 17.435 44.973

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 139.454237 7.167164 19.457 <2e-16 \*\*\*

Income 2.063285 0.041128 50.168 <2e-16 \*\*\*

Age 0.130534 0.074133 1.761 0.0791 .

StudentYes -97.955323 4.361103 -22.461 <2e-16 \*\*\*

Education 0.080938 0.401622 0.202 0.8404

Balance 0.233895 0.003208 72.901 <2e-16 \*\*\*

MarriedYes 3.262077 2.583521 1.263 0.2075

---

Residual standard error: 24.95 on 393 degrees of freedom

Multiple R-squared: 0.9744, Adjusted R-squared: 0.974

F-statistic: 2492 on 6 and 393 DF, p-value: < 2.2e-16

**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:**

Call:

lm(formula = Rating ~ Income + Student + Balance + Ethnicity +

Gender, data = CreditRating)

Residuals:

Min 1Q Median 3Q Max

-79.913 -11.709 3.447 17.673 47.352

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 153.55235 3.36437 45.641 <2e-16 \*\*\*

Income 2.07830 0.04034 51.516 <2e-16 \*\*\*

StudentYes -98.36044 4.35147 -22.604 <2e-16 \*\*\*

Balance 0.23328 0.00320 72.911 <2e-16 \*\*\*

EthnicityCaucasian -2.36824 3.07377 -0.770 0.441

EthnicityHispanic -5.52200 3.52754 -1.565 0.118

GenderMale -3.06521 2.50326 -1.224 0.222

---

Residual standard error: 24.97 on 393 degrees of freedom

Multiple R-squared: 0.9744, Adjusted R-squared: 0.974

F-statistic: 2488 on 6 and 393 DF, p-value: < 2.2e-16

**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.

Call:

lm(formula = Rating ~ Income + Student + Balance + Ethnicity +

Gender + Balance \* Ethnicity + Balance \* Gender, data = CreditRating)

Residuals:

Min 1Q Median 3Q Max

-82.711 -11.309 3.238 16.800 48.368

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 158.602997 4.223258 37.555 <2e-16 \*\*\*

Income 2.086134 0.040607 51.373 <2e-16 \*\*\*

StudentYes -98.000669 4.372489 -22.413 <2e-16 \*\*\*

Balance 0.222692 0.006214 35.839 <2e-16 \*\*\*

EthnicityCaucasian -8.240285 4.646746 -1.773 0.0770 .

EthnicityHispanic -10.653799 5.194322 -2.051 0.0409 \*

GenderMale -5.439409 3.781853 -1.438 0.1512

Balance:EthnicityCaucasian 0.011308 0.006656 1.699 0.0901 .

Balance:EthnicityHispanic 0.010129 0.007362 1.376 0.1696

Balance:GenderMale 0.004885 0.005488 0.890 0.3740

---

Residual standard error: 24.94 on 390 degrees of freedom

Multiple R-squared: 0.9746, Adjusted R-squared: 0.974

F-statistic: 1664 on 9 and 390 DF, p-value: < 2.2e-16

**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)

In all three models the baseline credit score has a range of ~19 when not accounting for any other predictors with a range between 139.4 – 158.6.

All three models suggest the strongest predictor of credit score is whether the person is a student. Being a student correlates to ~98 point drop in credit score in all models.

Other statistically significant predictors with relatively large Beta coefficients include Income and Balance. The effect of income was ~2.07points/1,000 dollars of additional income. The effect of Balance was ~0.2 points/dollar of additional balance. This is interesting as I would expect a larger outstanding balance to have a negative effect on credit score as noted in the predictor table

Models 2 and 3 explored the possibility of credit score bias based on ethnicity and gender. Model 2 did not produce statistically significant values for ethnicity and gender bias. The Beta coefficients for ethnicity and gender both included zero within one standard deviation from the mean beta coefficient indicating that we cannot reject the null hypothesis that there is no effect on credit score based on gender and ethnicity

Model 3 explores whether there is a significant interaction between balance and ethnicity/gender. The output indicates that the interactions for both ethnicity and gender are not significant. However, this model suggests there is a statistically significant effect of ethnicity on credit score, suggesting that Hispanics have a lower credit score than African Americans by approximately 10.7 points. The effect of Caucasian ethnicity on credit score is not strictly statistically significant based on a 5% error threshold, but is quite close at ~7.7% which may indicate the relationship is worth exploring further. Based on this model, Caucasians have a lower credit score than African Americans by ~8.2 points.