**MODULE 2**

**PREDICTIVE ANALYTICS**

**ANALYSIS BY CUSTOMER ID**

To build a suitable regression model, the annual salary derived for one hundred customers was used as the response variable. Age and gender were also taken from the existing data as well as a derived column called pay\_freq. The column pay\_freq, is a record of how frequently each customer is paid in a calendar year. The columns age, gender and pay\_freq were used as predictors for various models.

An analysis of the annual salaries showed that the values are not normally distributed. Therefore, a log transformation was used to approximate a normal distribution.

A screenshot of a cell phone

Description automatically generated

No significant correlations where found between the variables.

A screenshot of a cell phone

Description automatically generated

**SIMPLE REGRESSION MODEL**

Each simple regression model tested performed poorly. The following example is age as a predictor for annual salary. No predictors tested were significant and accounted for very little variation in the model.

lm(formula = transform\_y$log\_annual\_salary ~ transform\_y$age)

Residuals:  
 Min 1Q Median 3Q Max   
 -36843 -17877 -7387 15009 70081

Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 70494.4 7282.7 9.680 6e-16 \*\*\*  
transform\_y$age -139.5 215.6 -0.647 0.519

---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24760 on 98 degrees of freedom  
Multiple R-squared: 0.004256, Adjusted R-squared: -0.005905   
F-statistic: 0.4189 on 1 and 98 DF, p-value: 0.519

ANZ should not use this model for segmenting customers into income brackets.

**DECISION TREE**

To predict annual salaries, a decision tree was created with the variables gender and pay\_freq used as predictors. Annual salaries were divided into 2 categorical groups of either annual salaries less than $60,000 or those more than $60,000. Note that there were no annual salaries that were exactly $60,000.

A picture containing clock

Description automatically generated

To test the accuracy of the dataset, the 100 customers were broken into training and testing datasets. The results for the 24 customers in the testing dataset can be seen in the following confusion matrix.

**Less Than 60K More Than 60K**

**Less Than 60K**  7 5

**More Than 60K** 7 5

As a result, the accuracy of the decision tree is only 50%. In other words, only 12 customers annual salaries were correctly identified as either being less than $60,000 (7 customers) or more than $60,000 (5 customers). This is no better than tossing a coin but is still more accurate than the linear regression model which was a very poor predictor of annual salaries.