Josh Kohn

346039837

**Conclusion**

After cleaning up the data, I made two types of predictive models:

1. A linear regression model for predicting annual spend of the client
2. A logistic (classification) model for predicting if the clients annual spend would be below or above average.

Linear Regression Model

Our model was able to make a prediction on the test data with a mean squared error of $155,931,085.85 and an r-squared error of 0.654. Given the range of MSE, the quality of the model seems relatively low, however the quality of a model based on MSE should not be assessed in isolation but compared to other MSE’s. Since the average annual spending is above $50,000, it makes sense that the MSE will be a very large number.

If we show a histogram of the difference between the actual value vs the predicted value here is what we see:

The strengths of a regression model are:

* Suitability for tasks where predicting specific quantities (like sales forecasts, temperature predictions) is crucial.
* Easy interpretability of coefficients, indicating the direction and magnitude of the relationship between predictors and the target variable.

On the other hand, the weaknesses are:

* It can be sensitive to outliers, which can disproportionately influence the model's predictions.
* Assumes a linear relationship between predictors and the target variable. If this assumption is violated, model performance can degrade significantly.

Insights:

A black screen with white numbers

Description automatically generated

Here we can see that each subcategory (different cards, different genders and different cities) has different weights assigned to them. This is significant because, for example, it suggests a built-in difference between clients who have a Mastercard vs Visa card. This is critical information to have because the company might use it to focus their marketing towards a specific sector of people.

Histogram of Errors:

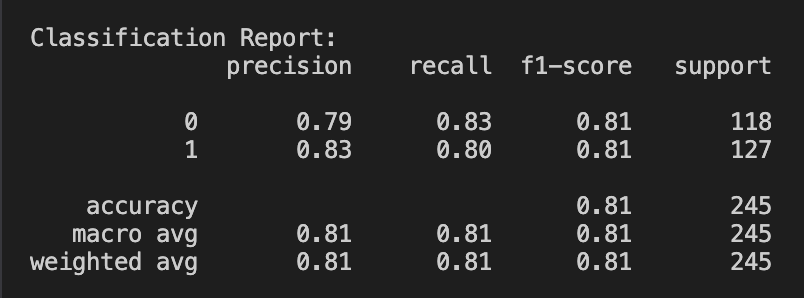
A blue graph with numbers

Description automatically generated

Here we see that most residual errors were +/- 10,000, which is overall a good sign for the model because it suggests that, even though not perfect, the model will more or less be accurate in determining the dependent variable (annual spending).

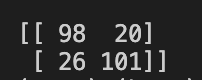
Classification Model:

For the classification model, I decided that an appropriate evaluation metric is a confusion matrix and an F1 score. This is because we want to see how many correct vs incorrect predictions the model made, and likewise the reliability (based on precision and recall) of the model. The results are as follows:



The precision, recall and f1 score all fall in the 79%-83%, which suggests a semi strong classification model.

The confusion matrix outcome is:



This suggest that the algorithm correctly reported 199/245 values, or 81%.

The strengths of a classification model are:

* Clear decision boundaries between classes, making them useful for tasks where categorizing data into discrete classes is needed.
* Classification models have well-defined evaluation metrics like accuracy, precision, recall, and F1-score, which provide clear measures of model performance.

On the other hand, the weaknesses are:

* Performance can suffer when classes in the dataset are highly imbalanced (i.e., one class dominates the dataset), leading to biased predictions towards the majority class.
* Many classification models assume linear relationships between features and class probabilities. They may struggle with datasets where the relationship is non-linear.
* Without separating the model to know the parameters (theta) assigned to each feature, it can’t be known what weights were given for each feature, since the model goes through a sigmoid function after the initial H(o) hypothesis function.