Angel Durrani

DATA 4950-D01, Spring 2023

Capstone Project

The data that is being used in this analysis is made up of information about each transaction from this particular online store. This analysis was conducted with these questions in mind: How does age affect the amount spent? How does gender affect the amount spent? The methods that will be used are linear regression and random forest.

Linear regression is one of many techniques that searches for relationships among variables (or columns) within a set of data. This is used for determining how some variables influence other or many other variables. The type of linear regression used in this analysis is the simple linear regression model, one independent variable and one dependent variable. I ran this model twice, once with Age being the independent variable and Amount Spent being the dependent variable, then the second time with Gender being the independent variable and Amount Spent again being the dependent variable. Linear regression calculates estimations of coefficients, which are then used as metrics to evaluate the model's performance.

Random forest is the second regression technique used; this is classified as a more complex model. The concept here is that many decision trees with different settings are combined, which reduces overfitting. This model randomizes the features given to the trees so that each feature, at some point, is factored in. The randomization technique that is used in the random forest model is bootstrapping. This technique creates new datasets from the original datasets that are the same size as the original, for each new decision tree, meaning the same row can be in a new dataset for a different tree. The next step in this model is called bagging, the aggregation of the results of the

trees. Each tree's results are accounted for and the prediction that results from these trees the most is the output for this model.

Simple linear regression was chosen because of its simplicity. When performance is measured, after running a linear regression model, the metrics are easily interpreted. The random forest model was selected because the same row can be used for each decision tree, taking account all features. Which could be beneficial when it comes down to bias, by not excluding features.

A linear regression model and a random forest model were run for age versus amount spent. Then another set of linear regression and random forest was run for gender versus amount spent. The performance metrics used in this analysis are R-squared, MSE (mean squared error), and MAE (mean absolute error).

R-squared is a measure of how much variation can be explained by the model. If the value of the R-squared is 1, that means every bit of the variation is explained by the model, so the closer to 1 that this value is the better. Both MAE and MSE are statistics that calculate error, so this would mean the closer to 0 the better, with 0 meaning no error at all.

	Age	**Gender**
Linear Regression	R-squared: - 0.01 MAE: 0.85 MSE: 0.97	R-squared: -0.01 MAE: 0.85 MSE: 0.97
Random Forest	R-squared: -0.08 MAE: 0.87 MSE: 1.04	R-squared: -0.0 MAE: 0.84 MSE: 0.97
Random Forest After Hyperparameter tuning	R-squared: -0.06 MAE: 0.86 MSE: 1.03	R-squared: -0.0 MAE: 0.84 MSE: 0.97

Neither of the models used performed well at all. None of the values are even close to what a good model should be, but if one had to be chosen, for gender the best performing and most explainable model would be Random Forest and for age, it would be linear regression. These are more explainable for the respective variables because of the better results. Another method, gradient boosting regression was attempted, but this also gave bad results.

Now, back to the questions that are to be answered by this analysis: How does age affect the amount spent? How does gender affect the amount spent? The answer to this would be there is no effect. Age and gender do not influence the amount spent. All three of these models are meant to analyze and report on whether there is a relationship between chosen variables and with these scores being so poor, the conclusion to this would be there is no relationship between age and amount spent or gender and amount spent.

The information derived from these models is useful in that with this particular store, they now know how to market to their customers. A reasonable assumption would be there are certain ways to advertise a product that would appeal more to a certain age group or to a certain gender, which is definitely something taken into consideration by marketers. Who are we selling to? But based on these models, there is no effect on the amount spent and gender or age. Products being sold here at this online store should be advertised with all age groups and genders in mind, since targeting a specific age group or gender would not have an effect on the amount being spent.

Background information resources:

https://realpython.com/linear-regression-in-python/#what-is-regression

https://data36.com/random-forest-in-python/