Customer Demographics Report

This document sets out to answer two questions, listed below. The answers to these questions are informed by applying Machine Learning concepts to the “Demographic Data” dataset:

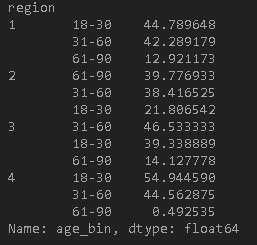
Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?

Is there any correlation between age of a customer and if the transaction was made online or in the store? Or do other factors correlate to an online or in-store transaction?

Analysis and Commentary

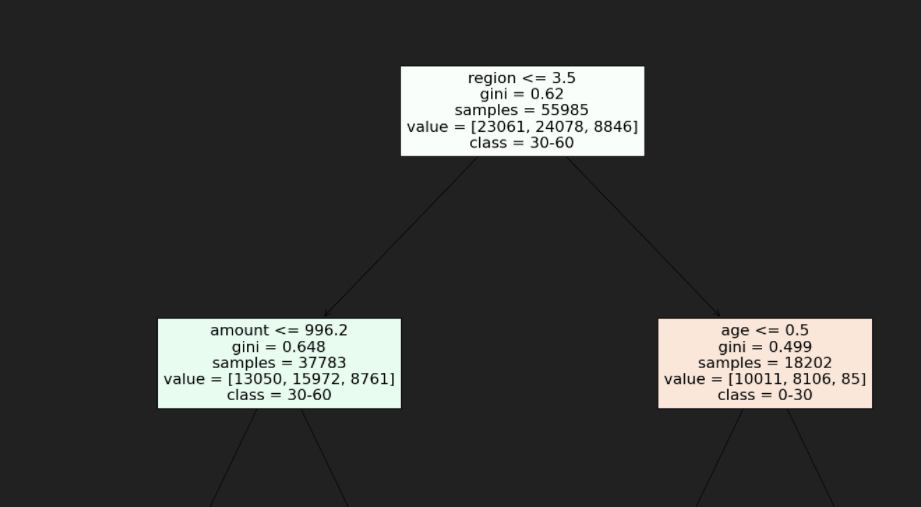
Three ML algorithms were used to answer the aforementioned questions. The algorithms were DecisionTreeClassifier, RandomForestClassifier and GradientBoostingClassifier. These algorithms can be used for classification-related problems, which is what this particular task involved. This task is a classification problem because we are trying to predict the occurrences of discrete, categorical variables. **An important consideration to make is that given that these algorithms are all related to classification problems, their relative performance should be very similar. In the case of this task, this consideration is true. Looking at the classification reports and the cross-validation scores, regardless of what the target variable was, all three algorithms had similar predictive results and performances. Most importantly, none of these algorithms were, or ever can be, “perfect” or “absolute”.**

That being said, where we see important differences is based on what target variable was used in ascertaining predictive power. For example, we can very clearly dismiss the predictive power of using the “items” variable. Looking at the classification report, we see a very low cumulative accuracy score (0.28 – 0.29) and low scores associated with each class (grouping) of items. This means that we cannot meaningfully glean information, such as trends, when looking at the distribution of “items”. The number of items in a purchase do not tell us much information about that purchase. We can’t generalize well.

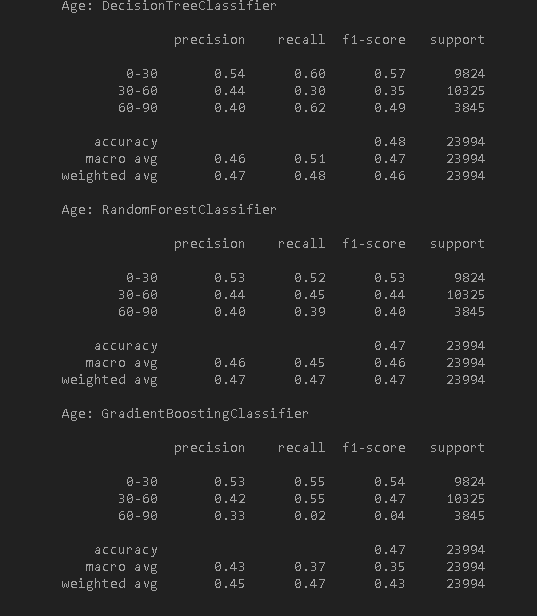
**There does appear to be differences in the age of customers between regions.** The overwhelming majority of purchases in the West region (most profitable region) are from people between the ages of 18 and 60. As for the least profitable region, the South region, the majority of purchases are from people between the ages of 31 and 90. This can be clearly seen from the following matrix which represents the age-group breakdown per region in the form of a percentage value: 

**There does not appear to be a correlation between age of customer and purchase type (online or in-store).** This is confirmed by looking at the correlation value (-0.17) and covariance value (-1.4) between the “age” variable and the “in-store” variable. What these scores express is that these two variables are weakly negatively related and have a weak negative effect on each other. In fact, no other variable is appears correlated (positively) with the “in-store” variable. That being said, only one variable appears to have a moderately clear negative covariance (-30.86) with the “in-store” variable, which is the “amount” variable. **Ultimately, it appears that the there are no clear trends or patterns related to purchase type. That is, whether or not a person will make a purchase online versus at a store does not seem discernible from the present composition of the data.**

To further reinforce my conclusions, relevant classification reports are presented below:



This is plot represents a Decision Tree when the target variable is “age”. As we can see, the Root Node is affected by the “region” variable. It is important to consider that this plot is not definitive and should be interpreted as “absolute truth”, as its results are contingent on the performance of the Decision Tree algorithm.



This represents the classification reports for the three algorithms used for this task, when the target variable is “age”. None of the algorithms appear particularly effective at predicting the age-grouping, although they all perform better when predicting the “0-30” age grouping and then struggle when predicting the “60-90” age grouping. While I am not certain, I imagine this has to do with the relatively low support (sample size, if I’m not mistaken) size associated with this age grouping.

Next Steps

Having said all of this**, I conclude my analysis by suggesting that more data be collected and expanding the number of features.** For example, additional features could include “marital status”, “educational level”, “income level” and so on. These potential features may elucidate potentially clearer trends within the data than the current features. Regardless, a larger dataset would certainly lend itself to containing potentially clearer trends than what may presently exist in the dataset now. Even if the prior claim ends up not being true, the algorithms would likely perform better if the data was larger because their support amounts would be larger. Either way**,** I believe the company should continue to focus its efforts in retaining its clientele in the West region as well consider optimizing the customer experience for the South region, as that region could potentially have greater sales**.** Ultimately**, I do not feel very confident in saying that clear and precise predictions can be made given the current composition of the data as well as due to the relatively poor performance of algorithms.**