Predicting Customer Income Based on Data

12.13.2021

—

Gavin South
CSE 450

Overview

From Kaggle, "Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers.

Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment."

From a further exploration of the data it's apparent that there are certain points of data associated with each customer to predict aspects of their life. In this situation we will use whatever is available to us in order to figure out how much a customer would make in a year, and thus how much they are likely to spend on online purchases. Understanding these trends can help tailor certain marketing objectives for specific people instead of guessing.

Goals

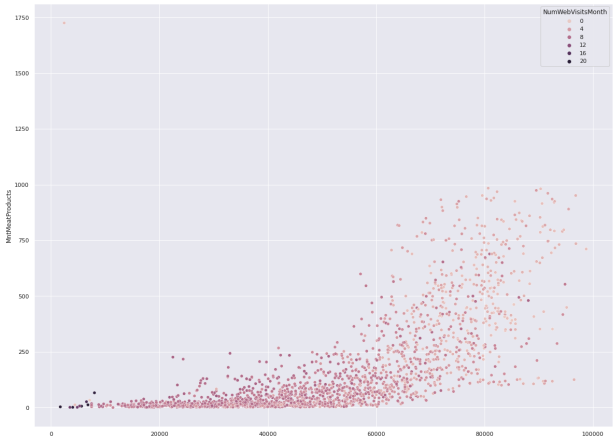
1. Explore the data using graphs and summaries for executive level decisions.
2. Build a machine learning model that can predict a customers income range based on consumer shopping behavior.
3. Make informed decisions and market to specific demographics within the database.

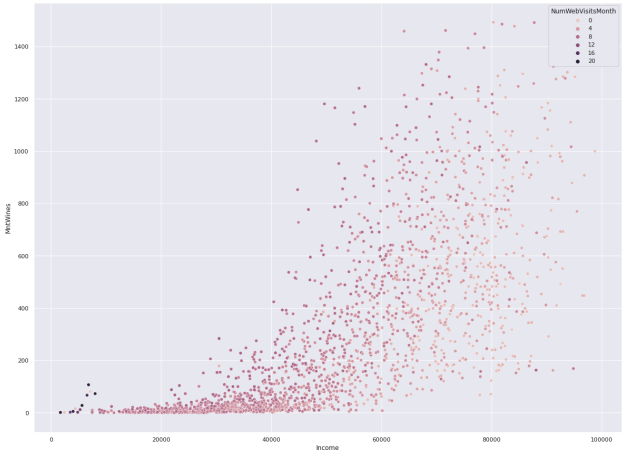
Data Exploration

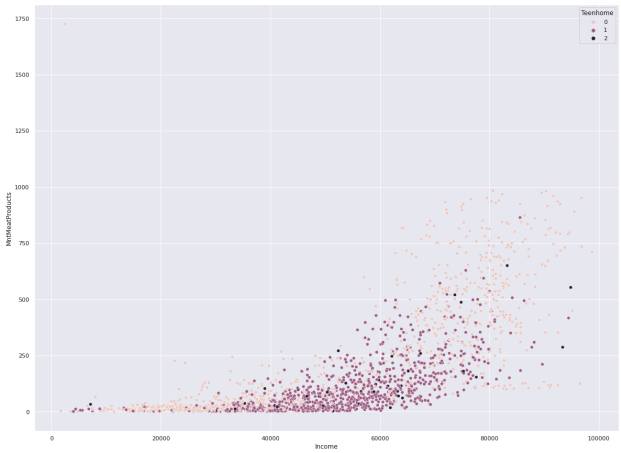
Building an XGBoost model seemed to be the best bet for finding connections and making predictions. Initially after throwing the data as it stood into the model, we found a R-squared value of .5 which suggests our model can explain about half of our data. But we figured this can be improved. After further inspection, there were some extreme outliers on income that skewed the model's performance drastically. We figured that as most of the customer base by proportion made under $100k a year, we cut off those extreme outliers and reevaluated our position. This resulted in a much better model that could describe some interesting actionable trends.

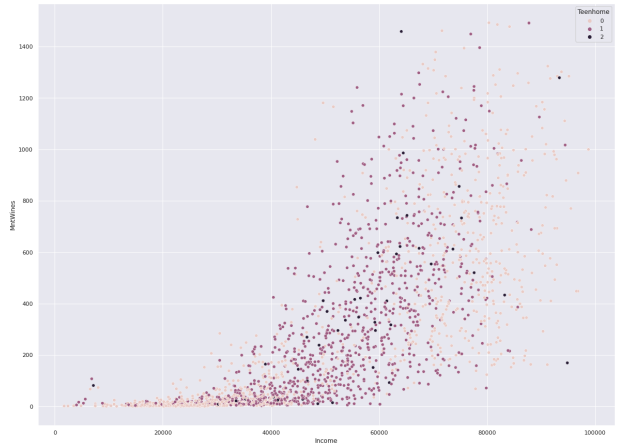
In these graphs are found correlations and actionable insights. We can see that the majority of large purchases are made by people that are making between $60k-$80k

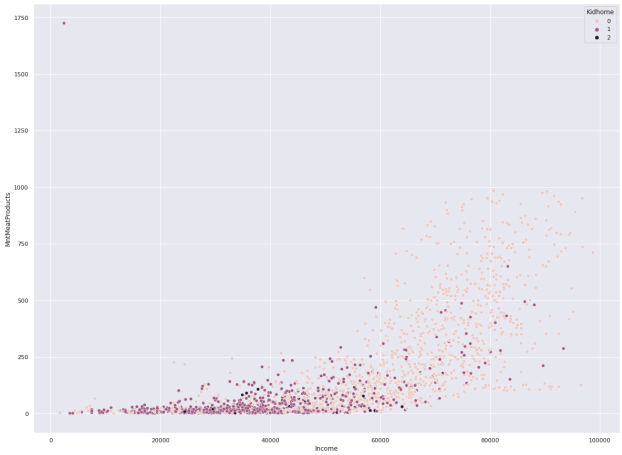
annually. Furthermore, and interestingly enough, people that visit the site less actually spend less and make less money. Among other things, alcohol is not bought in homes that have children around. Also makes sense but gives us some actionable insights.

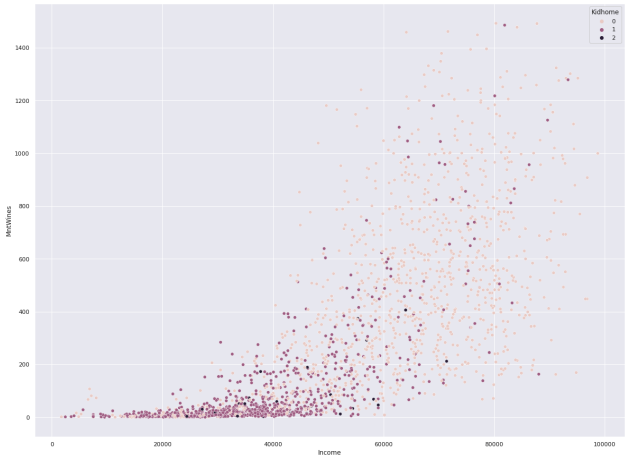










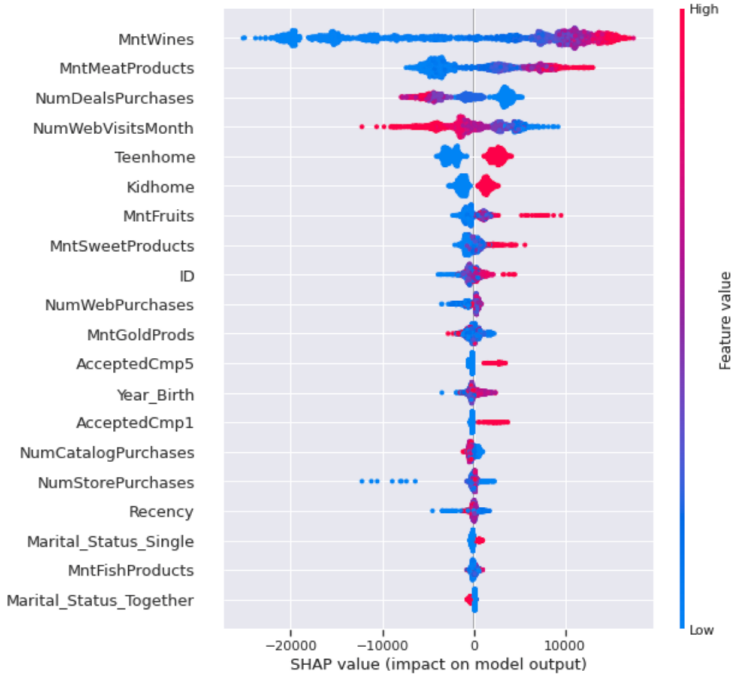


Modeling and Insight

We found again that a XGBoost was a very capable way of performing a regression on this set of data. We found that we could predict someone's income with the following levels of performance metrics:

```
RMSE: 7479.935330

R2: 0.869094
```



This is great news and shows that we have an accurate model that can guess a customer's income (and thus probably spending habits within the average range of $7.5k) . We also found that our model can account for 87% of all the data points in the set. It will likely continue to do so unless something drastic changes.

We found that there are a few different types of people that are shopping more from this site. People that spend a lot on wine, people that spend a lot on meat, and people that shop when there are deals around. We found that people who shop for deals more often are making less than $60k annually. People that buy wine a lot are making more than $40 annually and have no children in their home. It appeared to be that people with more financial responsibility (with children due to age or circumstances) did not spend more money. But, for the person that is making money and spending a lot and is free of those extra financial responsibilities, they have no trouble spending A LOT more money on average.

This information will help you all make more informed decisions and thus increase company income. This is the benefit of machine learning and predicting customer personality type, income, and demographics. We can offer services to more people and increase our own income threshold.

Google-Colab Work

https://colab.research.google.com/drive/1dhCqGCGQcbbPv_rZyrazT-kT2i2WSVwz?usp=sharing