Santander Product Recommendation

Drew Harmon

Purpose

My objective with this project is to create a recommendation model that can recommend items to users who are similar

I will be doing this with the collaborative filtering technique and utilizing cosine and pearson similarity to identify those users who are similar.

Data Wrangling/Cleaning

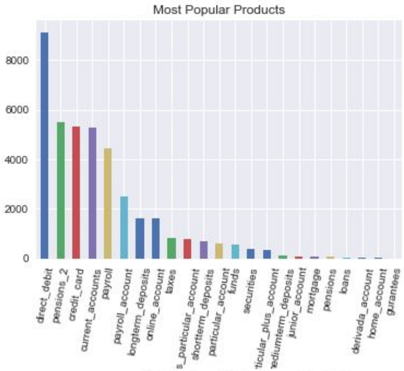
The data was provided by Kaggle, so it came pretty clean as it was.

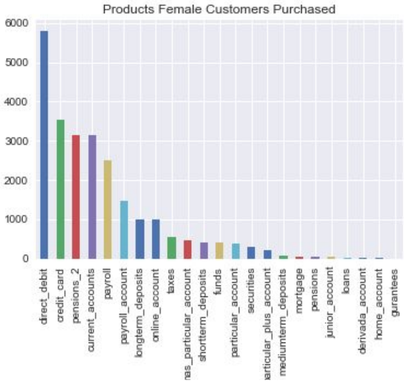
However, I did have to prepare the data for analysis by deciding on how to deal with NaN values, translating column names to English from Spanish, and changing the object type of most of the columns (Just as an example

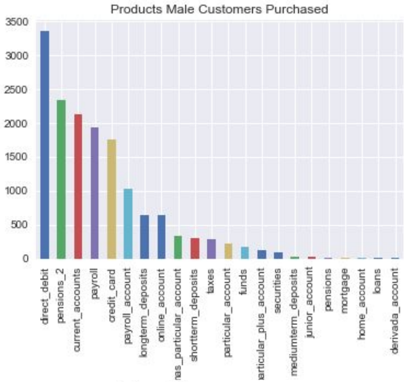
The most intricate part of the process was taking out values that didn’t belong. For example, if the only values that needed to be in a column were 1 and 0, I needed to take out random strings or other values that did not belong.

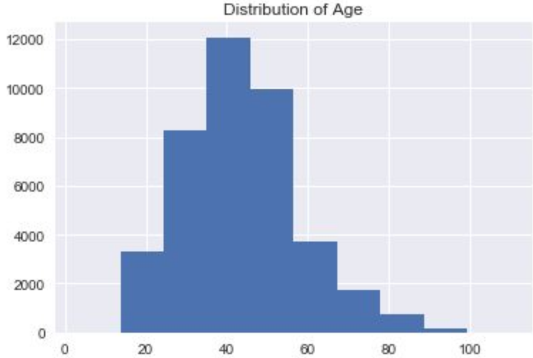
Interesting findings

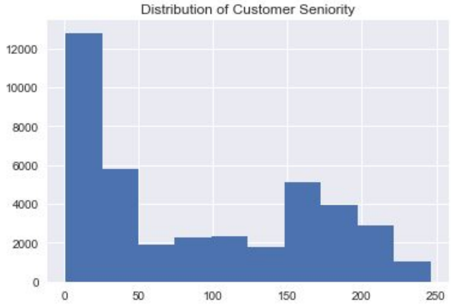
Exploratory Data Analysis

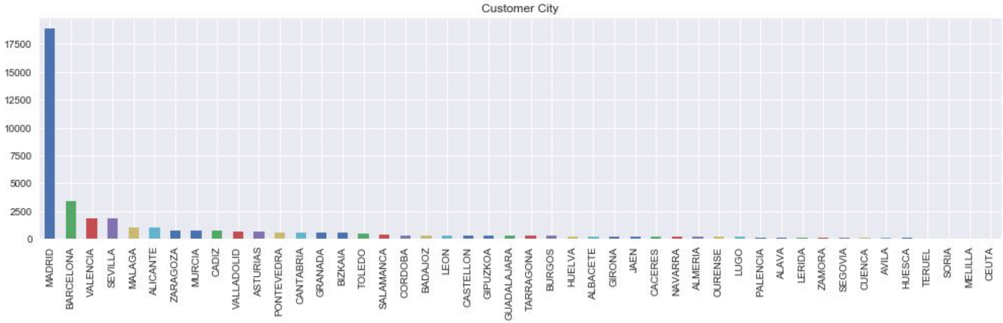

Most Popular Products

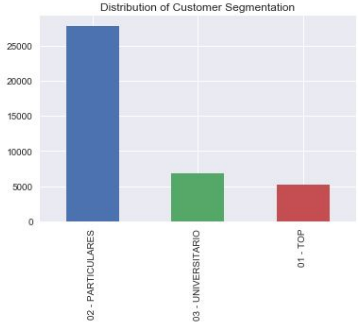

Products Female Customers Purchased


Products Male Customers Purchased


Distribution of Age

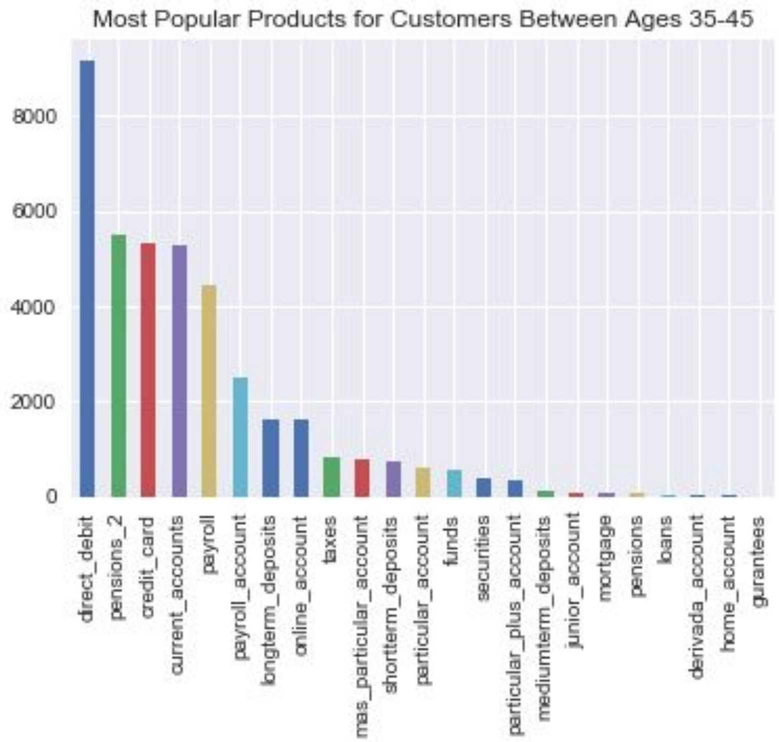

Distribution of Customer Seniority


Customer City


Distribution of Customer Segmentation

Most popular products by customers ages 35-45

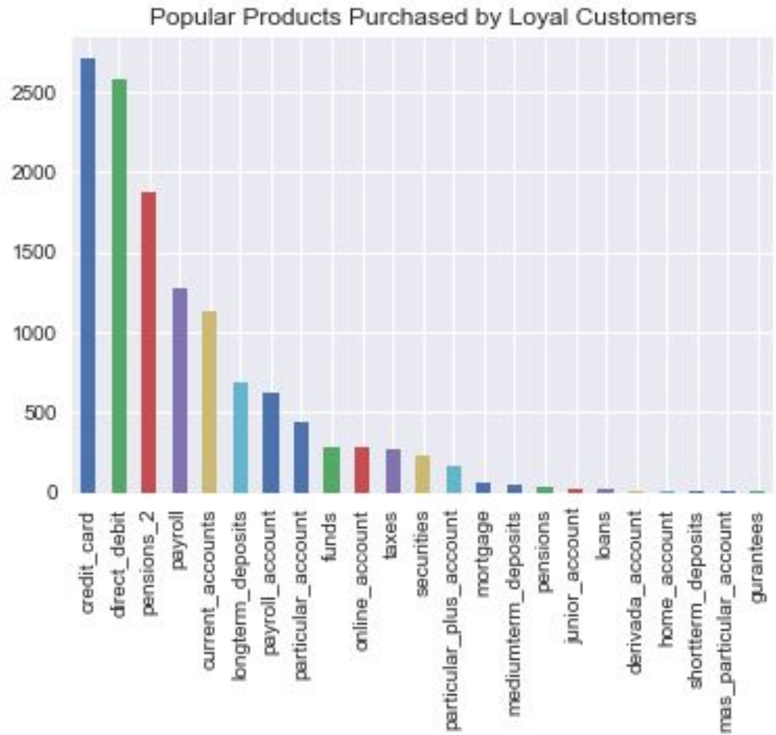
1. Direct Debit
2. Pensions
3. Credit Card
4. Current Accounts
5. Payroll



Most popular products for ‘loyal’ customers

1. Credit Card
2. Direct Debit
3. Pensions
4. Payroll
5. Current Accounts

Notice - The difference between the #1 and #2 products is very slim. This segmentation seems to prefer credit card and direct debit accounts almost identically.



Machine Learning

For the machine learning portion, I had to create a recommendation engine using the Turi Create package. I created 3 new datasets that would be easier for the model to ingest. I then decided to define similarity between users using two techniques - the cosine and pearson similarity methods.

After running those models, I evaluated them using the RMSE and precision/recall.

Results

I decided that either the pearson similarity model using the purchase dummies dataset OR the cosine similarity model using the purchase dummies dataset would both be appropriate as they were both scored almost identically by the RMSE and the precision/recall techniques outlined previously.

There were some popularity models that I tested out that actually scored better than these two, but my biggest objective was coming up with a model that has the ability to recommend products based on user similarity. Therefore, I decided to ignore them and pick either the models discussed above.