Simon Leahy

Capstone Milestone 2

Springboard Data Science Career Track

9/29/2020

Capstone Project 2

Marketing analytics is an extremely important facet of a company’s growth. The emergence of the online marketplace has drastically increased the amount of collectable data and potential actionable insight that can be extracted from that data. Data such as customer demographics, reviews, social media sentiment, geographic economic trends, and emerging elements of popular culture can all be extracted, collected, and analyzed through connections to online marketplaces.

Data

I am looking into a publicly distributed dataset from a Brazillan E-commerce Curator, OList. OList has created a network through which small and medium-sized businesses can advertise and distribute on an international scale. They provide an advertisement platform and logistics solutions to these companies to help them reach a broader market. In addition to the revenue they gather as a function of percentage of client sales, they have amassed a wealth of data that can be used to predict emerging trends and identify features that contribute to success in the marketplace. The dataset that they have shared features information at the product, client, end customer, geographic, and transaction level, exposing sales pipelines for products and industries that have experienced disproportionate growth and the potential variables that contribute to that growth. The data is located in two separate repositories, with roughly eight tables each in a relational database structure.

Problem

The existing problem is that e-commerce is a competitive field in which it is difficult to establish proprietary value. I believe there are patterns in customer behavior and features of customers, products, and circumstances that contribute to company and industry growth, and this information can be used to optimally target potential customers. For example, we might identify that a customer who purchases a type of product is likely to have demand for other specific product categories. This information, along with other analyses, would allow a company to optimize targeted marketing toward those customers. These patterns, once identified, can be used to allocate attention to those features and predict industries that are positioned for high economic opportunity.

Client

My client is Olist, and I will provide them with insights that they can sell to their clients along with information that they can use to give them a deeper understanding of customer dynamics within industry growth at a macro level The approach to solve this problem will involve the formatting of a broad database schema to extract data pertinent to applicable machine learning models.

Envisioned Data Science Methods that will be Considered

There will be opportunity to use a variety of data science methods as described in the table below, including linear and logistic regression, on contributors to industry growth, natural language processing transaction-level reviews, and clustering models on customer segmentation based on consumer habits. Specific opportunities include the use of NLP with Naive Bayes to create supervised models with review text and product ratings per transaction, clustering on our customer database pertaining to ordering habits, and Naive Bayes on a list of marketing qualified leads (MQLs) that result in deal closure. To clarify, marketing qualified leads are potential clients that Olist can take on and provide services to, and closures are MQLs that result in a company subscribing to Olist’s services. Each MQL has information on lead origin, which can be used as a qualitative variable to predict probability of closure. Once the lead has been closed, information is collected on the company, such as company behaviour, company type, business segment, and company maturity. These are all categorical variables that can be used to predict revenue the company will generate on Olist.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Business Question | Sample methods to use | Problem Type |
| 1 | Extract importance of features that contribute to company revenue | Linear Regression,  Tree-based Regressors | Supervised Regression |
| 2 | Extract importance of features that contribute to company growth, where companies are classified according to classes of growth, to be determined | Logistic Regression, Naive Bayes, Tree-based Classifiers | Supervised Classification |
| 3 | Sentiment analysis on reviews and ratings | NLP for sentiment analysis, Naive Bayes, Tree-based Classifiers | Supervised Classification |
| 4 | Segmenting customers by purchasing behavior | PCA, K-means. TSNE | Clustering |
| 5 | Predicting probability of closing MQLs using components of company | Logistic Regression, Naive Bayes, Tree-based Classifiers | Supervised Classification |

Deliverables

The deliverables for this project will be insights gathered from the business problems and methods above. An example of the structure of possible insights for each model is provided below:

|  |  |
| --- | --- |
| 1 | Business segment and city of origin contribute most to company weekly revenue |
| 2 | Companies with a shorter time between contact MQL and closure have a faster rate of revenue per week growth |
| 3 | With this model, we can predict satisfaction level based on review through sentiment analysis and naive bayes. |
| 4 | Customers who purchase products in the bed/bath category also purchase products in the health/beauty category |
| 5 | The strongest decrease in marginal probability of converting MQL to a closed account occurs 25 days after first contact |

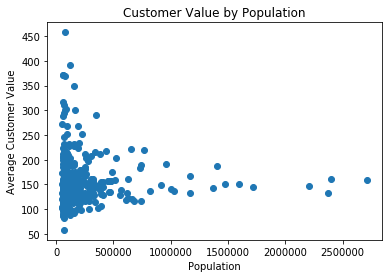
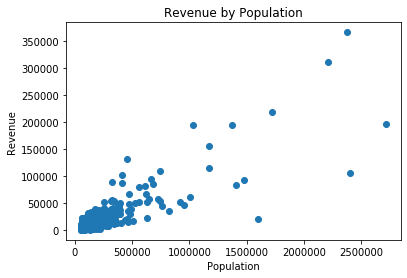
The output will be refined models within Jupyter notebooks, along with documentation and PowerPoint Presentation/Text document containing extracted insights.

The data wrangling requirements to complete this project involve creating data relationships to convert what currently exists in a database format into a tabular format that is fit for application of statistical analysis and machine learning models. There are two databases, one with nine and one with two tables, and all of these tables contain a unique primary key, and an external dataset that provides population information at the city level. It is essential to understand the structure of each table that makes up these databases and the relationships these tables have with one another. The shape and head function are used to look into each table, identify which tables relate to one another, and identify which table provides the most granular level of data. This is the set that will presumably be used for the application of statistical analysis and machine learning.

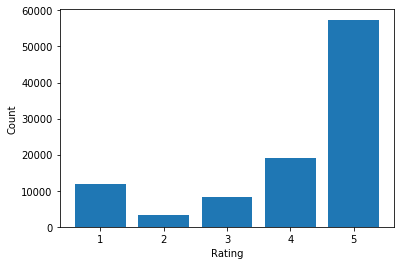
Once the tables are in a format to be partially viewed, they must be joined via pandas merge function to include all attach all extraneous data to the more finite levels that will be used for analysis, which is items. Through the methods below, all information associated with every item purchased through Olist can be viewed in one table.

This yields the Purchase\_Data table, which holds all items purchased and all possible respective data for each item purchased, product, price, shipping date, shipping price, product category, company that makes the product, pertinent information on the company such as industry, company location, date of joining Olist, and customer information such as customer zip code, city, and state. Identifying unique customers to the products they purchased will be especially helpful when performing customer segmentation. The merge block also attaches the external population information based on corresponding cities. This data table can be used to perform a pivot function on any variables associated with an ordered item.

One such pivot is revenue, unique customers, average value per customer, and population by city. This view can be used to understand the degree to which higher population affects the other business metrics. The plots below show relationship between both revenue and average customer value against population



The pearson R correlation between revenue and populations is extremely high (.95), indicating a strong relationship, which is somewhat intuitive. There is no relationship between average customer value and population. The rating distribution is also valuable information that can be extracted from a pivot of the Purchase\_Data table. The results are shown below.



review\_score order\_id Percentage

0 1 11858 11.8%

1 2 3235 3.2%

2 3 8287 8.3%

3 4 19200 19.2%

4 5 57420 57.3%

The majority of reviews are 5 and 4, indicating general satisfaction, but roughly 12% are poor reviews. The product categories with the

Top 5 Products (average weighting)

Count Average Review Score

product\_category\_name\_english

cds\_dvds\_musicals 14 4.64

fashion\_childrens\_clothes 8 4.50

books\_general\_interest 565 4.43

books\_imported 62 4.41

books\_technical 272 4.34

Bottom 5 products (average weighting)

Count Average Review Score

product\_category\_name\_english

fashion\_male\_clothing 145 3.53

office\_furniture 1788 3.51

home\_comfort\_2 31 3.38

diapers\_and\_hygiene 39 3.25

security\_and\_services 2 2.50

Top 5 revenue driving product categories

Total Revenue

product\_category\_name\_english

health\_beauty 1297761.15

watches\_gifts 1254577.20

bed\_bath\_table 1093842.41

sports\_leisure 1024909.58

computers\_accessories 942487.27

Bottom 5 revenue driving product categories

Total Revenue

product\_category\_name\_english

flowers 1110.04

home\_comfort\_2 773.17

cds\_dvds\_musicals 730.00

fashion\_childrens\_clothes 569.85

security\_and\_services 283.29

Another series of information we can extract from our current database is customer-level trends that can aid in customer acquisition and retention strategies. The return customer tag can be generated through below code:

Return\_Customers=pd.DataFrame(Purchase\_Data.groupby(['customer\_id'])[['order\_item\_id']].count())

Repeat\_Customer=[]

for i in Return\_Customers.order\_item\_id:

if i >1:

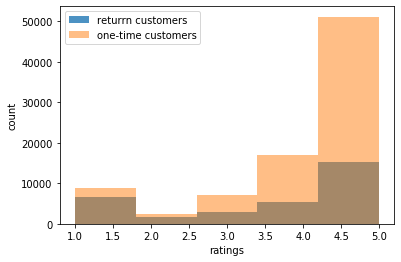
Repeat\_Customer.append(1)

else:

Repeat\_Customer.append(0)

Return\_Customers['Repeat\_Customer']=Repeat\_Customer

This view of data can be used to generate a look at average ratings for return and one-time customers. The below chart



Average return customer product rating is 3.65

Average one-time customer product rating 4.15

20.94% of return customer purchases were rated as 1

10.14% of one-time customers purchases were rated as 1

Growth Accounting-

# new customer

Churn rate

Resurrected

Retention

One vs many