Simon Leahy

Capstone Project 2

9/9/2020

Springboard Data Science Career Track

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.