BSTAT 625 Final Project Report

Ben Brennan, Mandy Meng Ni Ho, & Tahmeed Tureen

HuiJiang625, LLC

Introduction

Background

The integration of Machine Learning and Natural Language Processing (NLP) has been growing tremendously in the modern era. Powerful and flexible computers have allowed researchers in the field of NLP to apply statisical methodologies to large, unstructured data in the form of texts to draw insights from texts or create software application. Successful NLP applications such as the autocorrect feature on messaging apps, search engine suggestions, and question answering machines (i.e. `Siri`, `Cortana` etc.) have increased the field's popularity. As a result, many companies have started to invest in NLP to analyze the large amounts of textual data at their disposal.

Motivation

In this project, we spin a scenario where we are Data Science consultants at a world renowed consultancy firm, **HuiJiang625 LLC**. The primary clients of this consultancy group are airline agencies (i.e. Delta, American Airlines etc.) who are looking to gather some insights from the flight reviews they receive from their customers. Essentially, these airline agencies would like to identify which of their customers are most likely to recommend their airline to a friend(s). Being able to do so would potentially help the marketing team at these agency optimize their business through targeted marketing. For example, if an airplane knows that Person B is very likely to recommend the

Dataset

Raw Data

The dataset for this project is provided by Skytrax, a United Kingdom based consulting group that specializes in airline and airport reviews. This dataset consists of 41,396 observations with attributes such as customer review regarding flight, customer's country of origin, customer's ratings (i.e. overall, seat comfort, cabin staff rating etc.), customer's seat type (i.e. Business, Economy etc.) and finally a variable that represents whether or not a customer had recommended the airline after their review. The customer recommendation (variable name : `recommended`) is a binary variable and is the primary outcome of interest in this project. As discussed earlier, the objective of this goal is to create a model that can predict whether a customer will or will not recommend an airline. Therefore, the attribute holding the customer reviews is the main dependent variable in our analyses (variable name : `content`).

Data Processing

Data preprocessing for this project required two steps. The first was to include meaningful covariates that were not included in content. In essence, we needed to answer the question: Is there anything outside of the review (`content`) that is significantly associated with whether or not an individual is going to recommend the airline? After answering this question, we are then able to understand how to process the non-review part of our data to make it efficient in our alogrithms.

To answer this question, we used simple logistic regression to assess significance of certain covariates on the reccomendation. First, we excluded all *rating* variables (i.e. overall rating. wi-fi rating, food rating) as we felt the project would be too easy if we included those variables. We mainly focused on country of interest, type of traveller and what class the traveller was traveling in. We found, in our data set, that solo travellers were more likely to recommend the airline - but there was also about 39,000 missing values for this covariate, so it was not included. We found that a person from the US, as opposed to someone not from the US, is much less likely to recommend an airline. Furthermore, we found that if a person was flying first class or business class, they were more likely to recommend the airline compared to someone flying economy. Thus, we extracted two bits of information from the attributes of the customer - a binary indicator of whether they were in business class or first class vs. not and a binary indicator of whether a passenger was from the US vs. not.

The second question we sought to structure our data in order to answer was: how does a customers' review affect their recommendation? This question requires us to process the `content` column of our dataset in a reasonable way. We did this using the `tidytext` package in `R`. Essentially, we split the content of each review up into singular words, removed words that were not associated with sentiments from the *bing* list of words (`tidytext::get_sentiments('bing')`) and then constructured a document-term matrix (DTM). This matrix contains columns that are 1 if a word is present and 0 if that word is not present. For instance, a customer with a review such as "This was a great flight" would be split into words, and then "great" would be the only word that existed in this review. This customer would then have a row of data in the DTM that consisted of zero for every other sentimental word that existed in other reviews, and a 1 in the column correspoding to 'great'. In the end, our document term matrix is a 41,117 by 3,513 matrix, using about 1.1 GB of memory.

By joining our attribute predictors to our DTM, we obtained a dataset that was appropriately structured such that we were able to explore and attempt to answer the questions above.

Methodology

To create our predictive model

Statistical Models

- **Naive Bayes**

Let R be the event that a customer recommends the airline.

$$P(\text{R } | \text{"}Great\ Time\text{"}) \propto P(\text{"}Great\ Time\text{"} \mid \text{R}) = P(\text{"}Great\text{"} \mid \text{R}) \times P(\text{"}Time\text{"} \mid \text{R})$$

- **Logistic Regression**
- **Logistic Regression (No Intercept)**
- **Support Vector Machine**
- **Random Forest**

This shit failed so we legit didn't use it.

Technology Stack

Results

Model	Time (s)	Recall	Precision	F1 Score
Naive Bayes	0	0	0	0
Logistic Regression	0	0	0	0
Logistic Reg. (No. Intercept)	0	0	0	0
SVM	0	0	0	0

Deliverables

Contributions

- Ben Brennan
- Mandy Meng Ni Ho
- Tahmeed Tureen

Repository

The work directory of this entire project has been published on a GitHub repository, which can be accessed via the following link: (https://github.com/benbren/airplanes)