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

Methodology

To create our predictive model

Statistical Models

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: