

Product Pricing Suggestion

Using ridge regression

Randy Lao | Capstone Project 2 | March 12, 2017

# TABLE OF CONTENT

## ABSTRACT

## Introduction

## objective

## Representing and mining text

## overview of ridge regression

## obtaining the data

## data preperation/cleaning

## exploratory data analysis

## modeling

## test evaluation

## CONCLUSIOn

## ABSTRACT

This paper will cover the usage of logistic regression as a modeling algorithm to predict employee attrition risk within a company based on employee data. In this project, I will be covering my analysis and approach through different process flows in the data science pipeline. The main goal is to understand the reasonings behind employee turnover and to come up with a model to classify an employee’s risk of attrition. A recommendation for a retention plan was created, which incorporates some best practices for employee retention at different risk levels of attrition.

## INTRODUCTION

Can you automatically suggest prices to online sellers? Product pricing gets even harder at scale, considering just how many products are sold online. Clothing has seasonal pricing trends and is heavily influenced by brand names, while electronics have fluctuating prices based on product specs. **Mercari**, Japan’s biggest community-powered shopping app, knows this problem deeply. They’d like to offer pricing suggestions to sellers, but this is tough because their sellers are enabled to put just about anything, or any bundle of things, on Mercari’s Marketplace.

## OBJECTIVE

Mercari wants to build an algorithm that automatically suggests the right product prices. This will help both the seller and buyer with their prices, and hopefully yield more transactions. Mercari will provide the user-inputted text descriptions of their products, including details like product category name, brand name, and item condition.

1. **Define the objective in business terms:** The objective is to come up with the right pricing algorithm that we can use as a pricing suggestion to the users.
2. **How will your solution be used?** Allowing the users to see a suggested price before purchasing or selling will hopefully allow more transaction within Mercari’s business.
3. **How should you frame this problem?** This problem can be solved using a supervised learning approach.
4. **How should performance be measured?** The evaluation metric for this regression problem should be RMSE (Root Mean Squared Error), but in this competition it’s evaluated on RMSLE (  
   Root Mean Squared Logistic Error).
5. **Are there any other data sets that you could use?** To get a more accurate understanding and prediction for this problem, a potential dataset that we can gather would be more user information. Features such as location, gender, login information could affect it.

## representing and mining text

Since, text is the most unstructured form of all the available data, various types of noise are present in it and the data is not readily analyzable without any pre-processing. The entire process of cleaning and normalization of text, making it noise-free and ready for analysis, is known as **text pre-processing**.

**Fundamental Concepts**

The importance of constructing mining-friendly data representations; Representation of text for data mining.

**Important Terminologies**

1. **Document:** One piece of text. It could be a sentence, a paragraph, or even a full page report.
2. **Tokens:** Also known as terms. It is simply just a word. So many tokens form a document.
3. **Corpus:** A collection of documents.
4. **Term Frequency (TF):** Measures how often a term is in a single document
5. **Inverse Document Frequency (IDF):** distribution of a term over a corpus

**Pre-Processing Techniques**

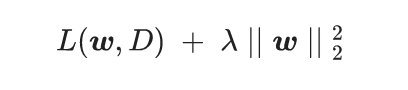
1. **Stop Word Removal:** stop words are terms that have little to no meaning in a given text. Think of it as the “noise” of data. Such terms include words, “the”, “a”, “an”, “to”, and etc..
2. **Bag of Words Representation:** treats each word as a feature of the document.
3. **TFIDF (Term Frequency Inverse Document Frequency):** a common value representation of terms. It boosts or weighs words that have low frequencies. For example, if the word “play” is common, then there is little to no importance. But if the word “mercari” is rare, then it has more weight/importance.
4. **N-Grams:** Sequences of adjacent words as terms. For example, since a word by itself may have little to no value, but if you were to put two words together and analyze it as a pair, then it might add more meaning.
5. **Stemming/Lemmatization:** converting a word into its common base form

## OVerview of ridge regression

Ridge Regression is a type of regression that use L2 Regularization, which aims for low training error while balancing against model complexity.

**L2 Regularization is about**:

* Penalizing really big weights
* Reducing model complexity
* Shrinking coefficients, but not eliminating them (unlike L1 Regularization)



**L2 Regularization Loss Function:**

* **L**: aims for low training error
* **Lambda:** A scalar value that controls how weights are balanced
* **W:** weights are balanced against complexity

## OBTAINING THE DATA

The data was found from “Mercari Price Suggestion Challenge” provided by Kaggle’s website: <https://www.kaggle.com/c/mercari-price-suggestion-challenge/data>

The most interesting part about working with this dataset is the fact that it’s huge. It contains about 2 million rows when combining both the train and test set. Because of this, I did some sampling of the dataset to perform my EDA a bit quicker. I’ll be using Python as my language of choice.

## DATA PREPERATION/CLEANING

Majority of the data preparation for this dataset will involve pre-processing of the text data. There are some missing values in the dataset as well and I imputed them with Nan’s for simplicity. Here are some of the pre-processing steps I did:

* 1. **Handling Missing Values –** Replaced the missing values with Nan.
  2. **Lemmatization** performed on item description
  3. **Label Encoding** – Turned categorical values into 0’s and 1’s
  4. **Tokenization** – Given a character sequence, tokenization is the task of chopping it up into pieces
  5. **Scaling** – Scaled the target variable (price)

**Data Column Description:**

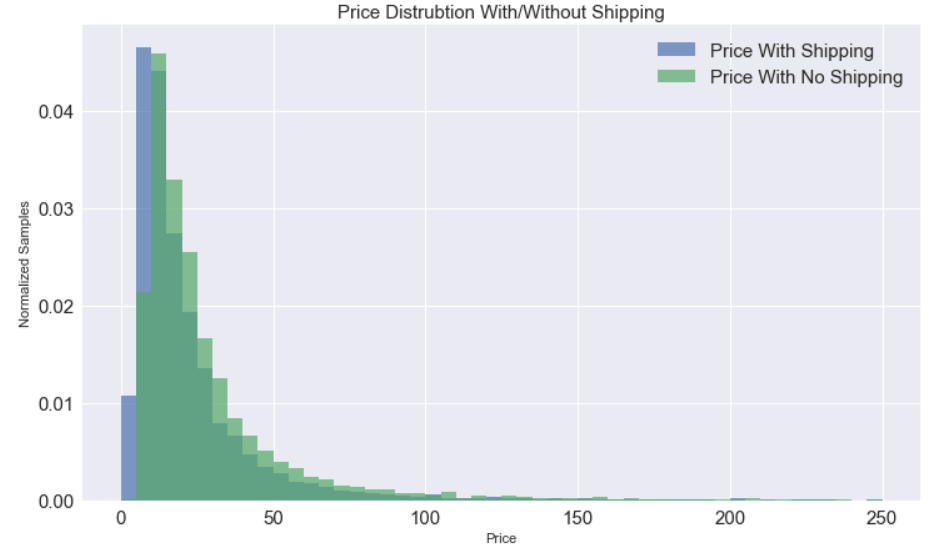
* 1. **train\_id / test\_id** – id of the listing
  2. **name –** the title of listing.
  3. **Item\_condition\_id** – the condition of the items by seller
  4. **Category\_name –** category of listing
  5. **Brand\_name –** brand of listing
  6. **Price –** the price for the item sold. Our Target variable we want to predict, which is in USD.
  7. **Shipping –** 1 if shipping fee paid by seller and 0 by buyer
  8. **Item\_description –** the full description of item.

## EXPLORATORY DATA ANALYSIS

**Price Distribution**

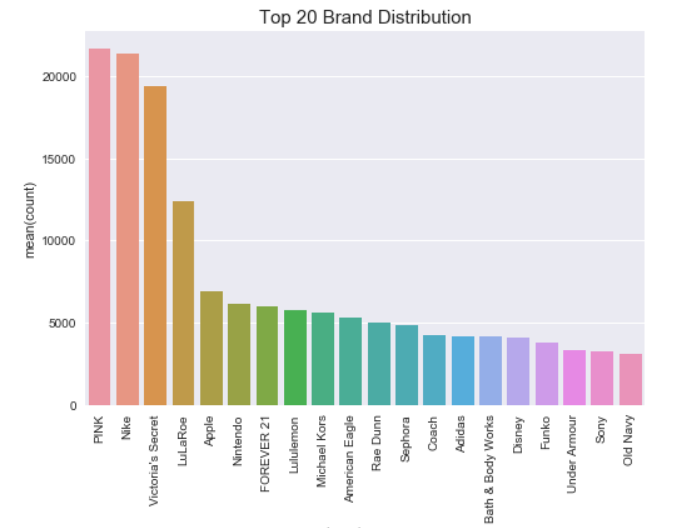
**Summary:**

* The mean price in the dataset is **26 Dollars**
* The median price in the dataset is **17 Dollars**
* The max price in the dataset is **2000 Dollars**
* Due to the skewed dataset, the **median** price is a more reliable price to gauge off of.



**Top 20 Brand Distribution**

**Summary:** Majority of the top brands are clothing brands and electronics. **PINK** and **Victoria Secret** are among the top 3 brands and are typically towards **female** customers.



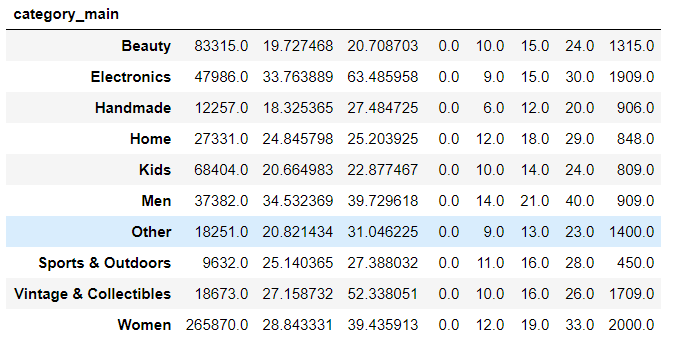
**Main Category**

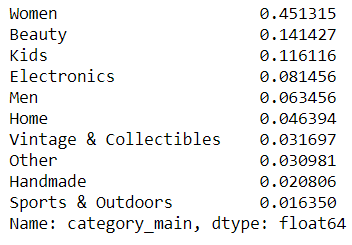
**Interesting findings:**

* Women and Beauty take up majority of the distribution
* Women and Beauty take up 56% of the distribution

**Questions to ask:**

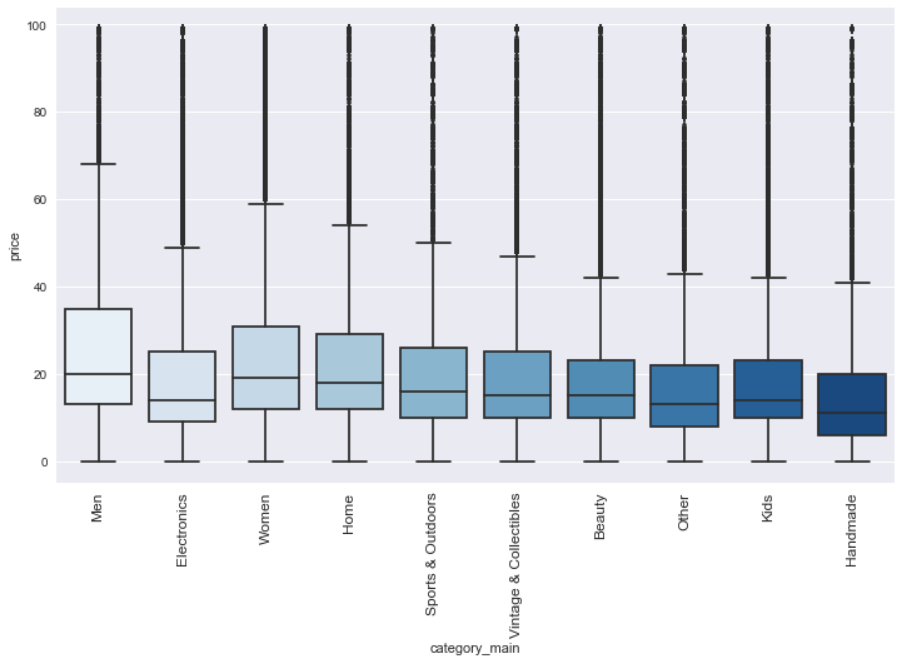
* Can we create a gender category (Female, Male, Nuetral). Example: Three categories means three gender types. If two of them are female, then we classify as a female purchaser. If two of them are male, then we classify as male. If male/female/neutral then?
* Does gender play a role in price?
* Can we create an age category?



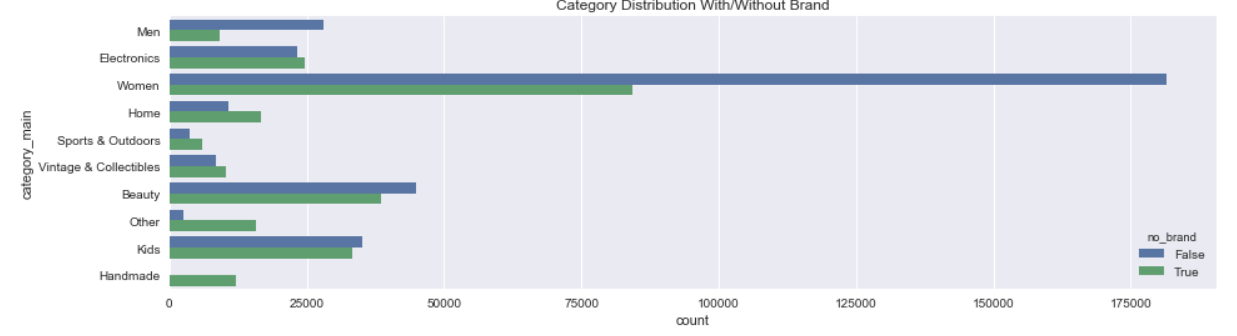


**Main Category Distribution (Boxplot)**

**Summary:** Looks like the prices are evenly distributed across categories. The **Men** category seems to be the only one that really averages out the most.

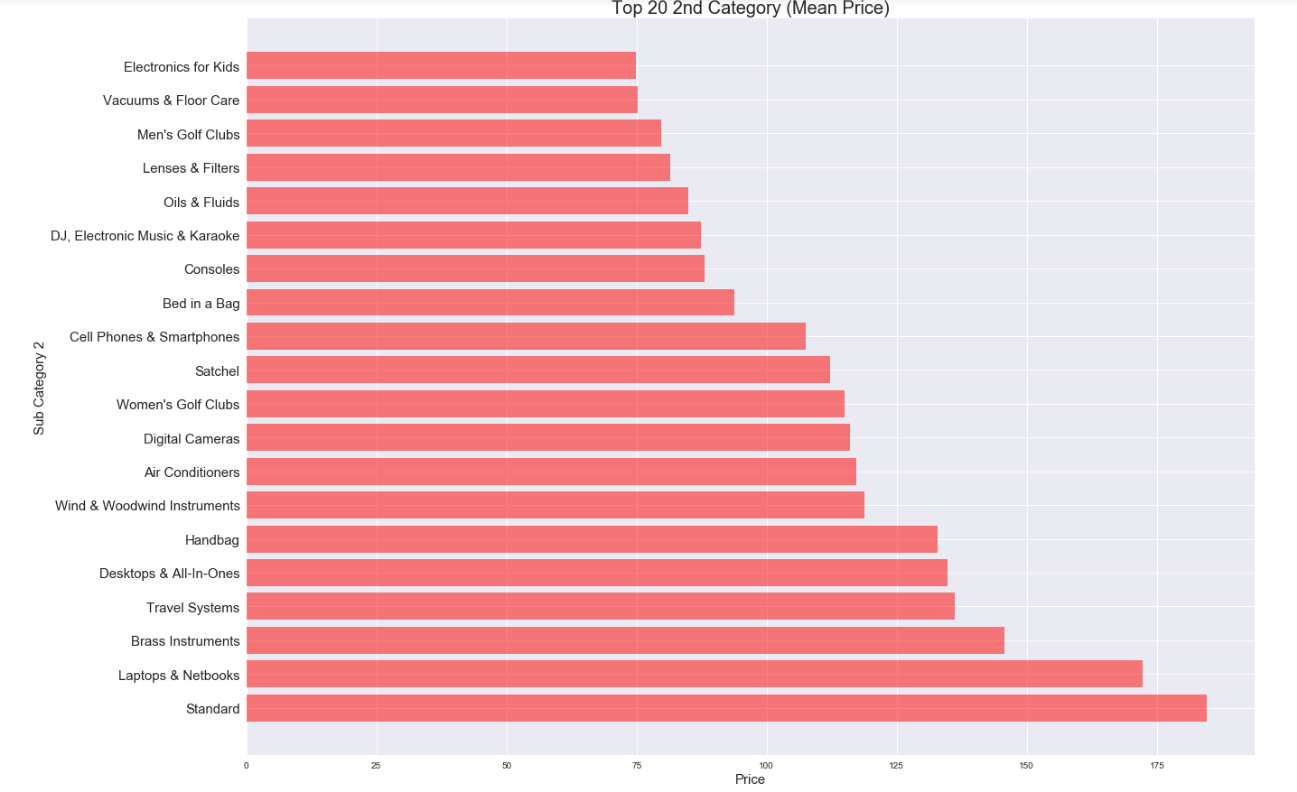
****

**Main Category With/Without Brands (Boxplot)**



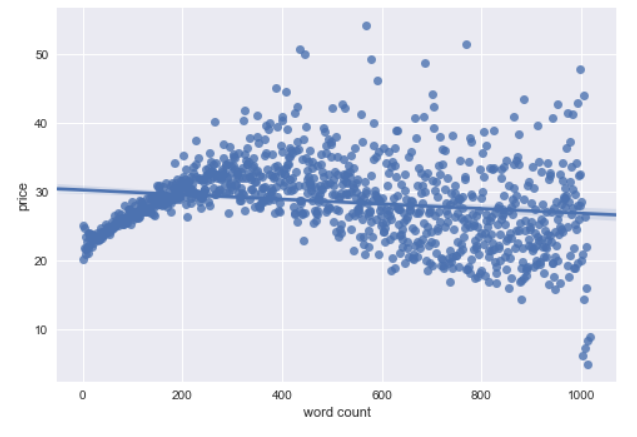
**Second Category Price Distribution:**

The majority of the second category falls under electronics and then by female products.

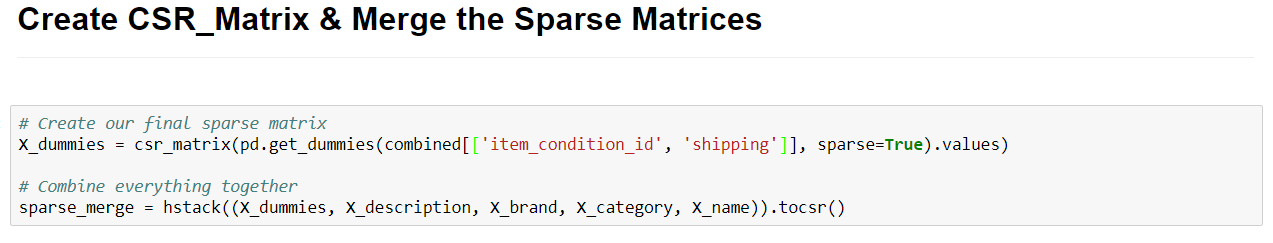


**Item Description Price VS Word Count**

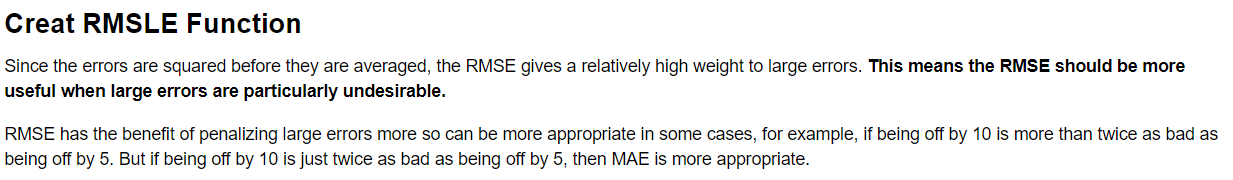
**Summary:** There is a positive linear relationship between word count and price from about 0-300 words. After that there is a gradual negative relationship, which drops at about the 1000 word point.



## Modeling



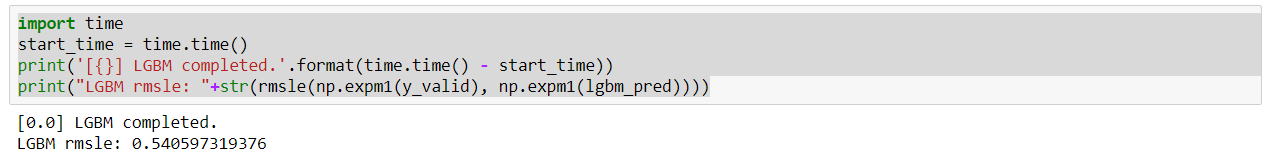
After getting all the variables cleaned and pre-processed, I had to convert them into a sparse matrix because there are thousands of word features. A CSR Matrix will allow better utilization of memory, efficient row slicing, and fast matrix vector products.



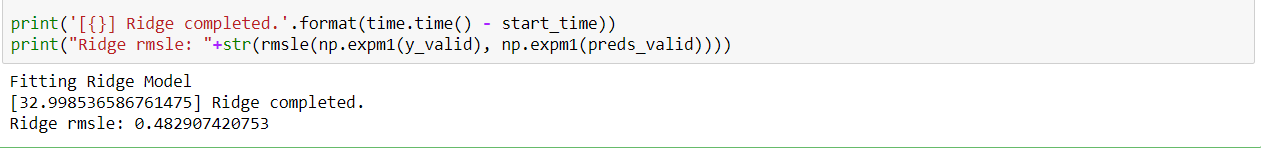
**Two Models Choices**

* 1. **Light Gradient Boosting Machine (LGBM) –** The reason why I used this algorithm is because it’s a good model to use on big data sets. It has fast training sped and high efficiency, low memory usage, good accuracy, and good compatibility with large datasets.
  2. **Ridge Regression –** As I explained before above, this algorithm is great off the shelf model to use on any regression task.

**LGBM RMSLE – 0.54**

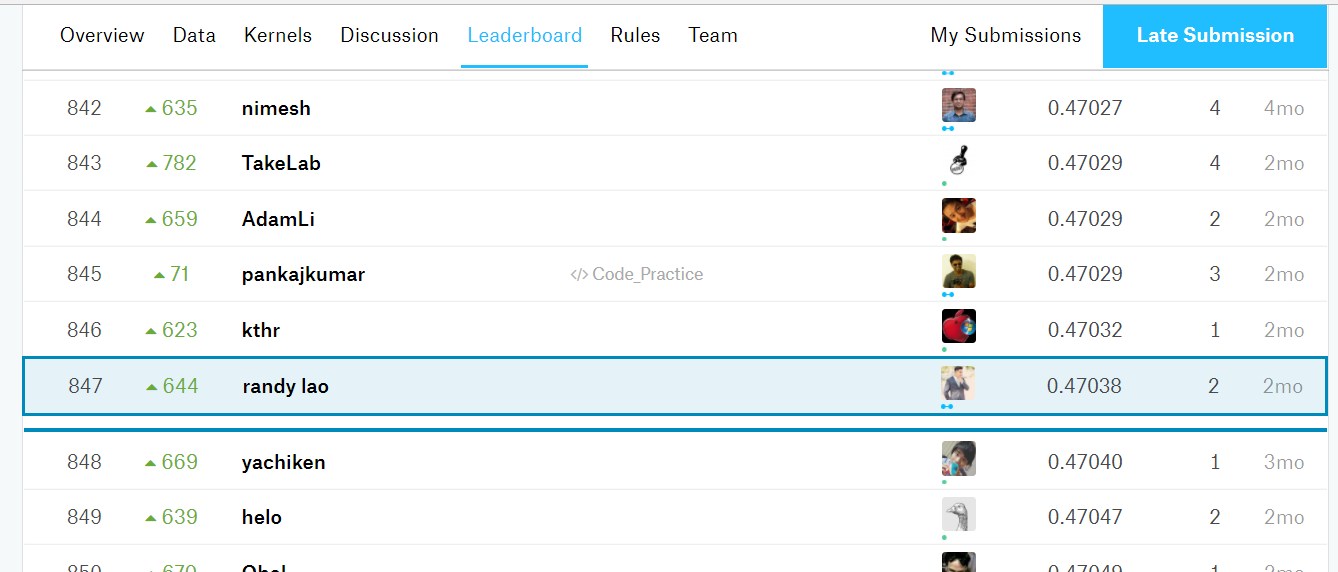
****

**Ridge Regression RMSLE – 0.48**

****

## Test evaluation

**Leaderboard Ranking (Top 36%) – RMSLE 0.47**

****

## conclusion

I am happy to have done this competition because it has opened up my mind into the realm of NLP and it showed me how much pre-processing steps are involved for text data. I learned the most common steps for text pre-processing and this allowed me to prepare myself for future work whenever I’m against text data again. Another concept that I really learned to value more is the choice of algorithms and how important computation is whenever you’re dealing with large datasets. It took me a couple of minutes to even perform some data visualizations and modeling. Text data is everywhere and it can get messy. Understanding the fundamentals on how to tackle these problems will definitely help me out in the future.