**Mercari Price Prediction with Natural Language Processing**

**--INFO7390 Final Project**

Jiali Cheng 001822320, Sicheng Zhang 0018

1. Introduction

In this project, we are dealing with Mercali price prediction, which is a Kaggle competition. Mercali is a C2C online trading platform where customers can post their products with some description and wait for the buyers. While setting a right price for a higher and quicker sale comes to the problem. In this process, Mercali, the platform, will help the seller set a proper price based on the features and descriptions provided by the seller. And we will use the Mercali sales data to build a machine learning model helping the platform achieve this goal. Then we will deploy this model on the cloud so that everyone can access this prediction model if they need with a REST API.

In the dataset, there are 6 common features and 2 columns of plain text. 6 features appear to have been a little less for an entrepreneurial prediction model with a huge amount of user and we will see later the score of this raw model is relatively low. So we leverage some NLP techniques to extract features from the text, including Bag-of-Words, Word Embedding and TD-IDF, and build a more powerful model.

For the prediction model, we use statistical method, ridge regression, random forest and neural network for prediction. And we compare they results both between various models and between whether using the text features or not.

1. Data Engineering
   1. Data Ingestion

We get this data from Kaggle in CSV format. We import it with Jupyter Notebootk.

* 1. Exploratory Data Analysis

After getting the data, we first look at its summary, containing object type and few values of each column.

We grouped the data in terms of state and calculate the total issued loan amount and number of issued loans within each state.

* 1. Data Cleaning and Preprocessing

We fill all the missing data with the string ‘missing’, which mean we take missing value as one of the categories. Then we encode the category data, which include the brand name and category of products.

1. Prediction without NLP
   1. Statistical Method

For this kind of data with a lot of category values, using group average first comes into our mind. We group the data based on the features of category, shipping information, decreases the data set we have to process.

* 1. Regression and Random Forest

1. Prediction with NLP

Now we will leverage the information given by the columns of ‘name’ and ‘item description’. We tried different ways. One is extracting informative features from the text. Another way is to embed the words into vectors.

* 1. Feature Extraction

First we define the features to be extracted as follow: item name, color, size, style, aging of the style, fabric, wash style, quality, sentiment of the description and . And we extract these features with regular expression.

* 1. Word Embedding

Word Embedding is a natural language processing model where word is represented as a vector. A sentence is the sum of all its words’ vector. We multiply the sum by a matrix and add some bias. Then we put the sentence vector into a neural network for training. There are different ways of embedding. We tried Bag of Words, Word to Vector and TF-IDF.

* + 1. Bag of Words

Bag of words counts the frequency of each word in the text and can be regarded as a one-hot embedding of words.

Since we are not focusing how to build a neural network, we use a neural network build by ‘noobhound’[[1]](#footnote-1)[1] from Kaggle.

* + 1. Word to Vector

Word to vector embed words in a more meaningful way. Words are transformed into vectors, or multi-dimensional spaces. And the direction of the vector the meaning of a word. For example, if we know ‘king’ points to ‘queen’, given the word ‘man’, we could get the vector of ‘woman’.

* + 1. TF-IDF

TF-IDF stands for Term Frequency – Inverse Document Frequency.

1. Deployment

We deploy the model on Google Cloud Platform with Flask. And everyone can access our prediction model with a web service at the IP address we gave.

Navigate to the website and you will see this UI where there are four demoes that you can run. They are all real data from the dataset. Or you can input any data and description to let our model predict the price of your product based on the information you provide. We will just show one of the demoes below.

1. Conclusion

Natural Language Processing is .

1. [1] Special acknowledgement to ‘noobhound’, a contributor of Kaggle. We use his neural network model for training the embedded vectors. [↑](#footnote-ref-1)