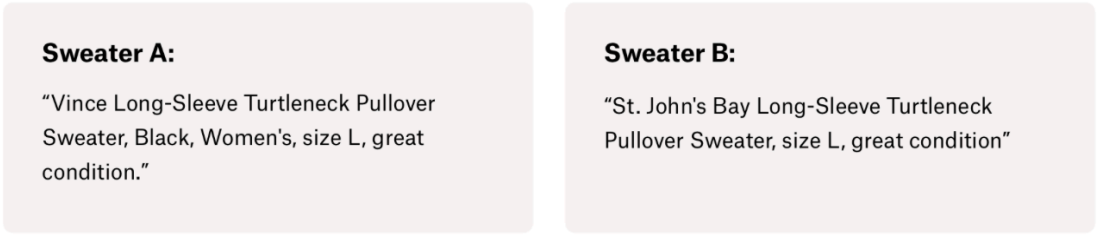
INFO7390 Advances in Data Science and Architecture Final Project Proposal
Kaggle Mercari Price Suggestion Challenge
Can you automatically suggest product prices to online sellers?

Part 1: Motivation

It can be hard to know how much something's really worth. Small details can mean big differences in pricing. For example, one of these sweaters cost $335 and the other cost $9.99. Can you guess which one's which?



An appropriate sales price setting is one of the most important part when companies do marketing research. Only if choosing the best sales price, the companies can gain the maximum profits. Different products in the same type, which normally depends on the brand, may facing different customers. Their purchasing power decide which basically price section your product should belong to, and then modify the price by the details of the product.

Product pricing gets even harder at scale, considering just how many products are sold online. Clothing has strong 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.

In this final project, we will build an algorithm that automatically suggests the right product prices. We will use the provided user-inputted text descriptions of their products, including details like product category name, brand name, and item condition.

Part 2: Data Set Introduction

In the final project, we are going to use Mercari's sales data set, which is the newest competition on kaggle, to predict sale price base on provided columns, including item name, item condition, category name, shipping fee and the most important part 'item description'. This data set has more than 1 million instances and 8 columns.

Part 3: Approach

3.1 EDA

- Do a detailed and in-depth EDA
- Summarize the insights of the feature item condition ID which has already been labeled.
- Summarize the insights of other.

3.2 Preprocessing

- Clean missing data
- Do other preprocessing if necessary

3.3 Clustering

We will try both creating one model for all products and segmenting data into clusters and then building prediction models specific to each cluster. We come up with three methods:

- No cluster
- Manually cluster the product into several clusters based on one feature
- Algorithmic clustering

3.4 Prediction

Our next goal is to predict the product prices.

Raw prediction:

- Given a piece of structured information of a product, predict the price
- Using linear regression, random forest, neural network or other models
- Compute the metrics including RMSE, MAE, MAPE and enhance the model

Advanced prediction:

- Us the item description provided, do some NLP on the item description and compare the performances of with and without doing NLP
- (Optional add-on, if possible) provide the figure of a product and extract the features

3.4 Deployment

- Deploy the best prediction model as web services
- Access the model with REST API

Part 4: Expected Results

The main purpose of this project is predicting the sales price. Namely, given some information of a product, including user-inputted text descriptions of their products, details like product category name, brand name, and item condition, the model will predict the price of the product. And as an improvement, we will introduce some approach to enhance the accuracy.

This prediction model will contribute to not only setting the price, but also play an important role in the enterprise decision making process.