Proposal: Mercari Product Price Suggestion

Product pricing gets 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 and they would like to offer pricing suggestions to sellers. This problem is challenging because sellers can put just about anything, or any bundle of things, on Mercari's marketplace.

The goal of the project is to build a model that automatically suggests the right product prices.

The data is obtained from https://www.kaggle.com/c/mercari-price-suggestion-challenge/data and consists of the columns:

- train_id or test_id - the id of the listing
- name - the title of the listing. Note that we have cleaned the data to remove text that look like prices (e.g. $20) to avoid leakage. These removed prices are represented as [rm]
- item_condition_id - the condition of the items provided by the seller
- category_name - category of the listing
- brand_name
- price - the price that the item was sold for. This is the target variable that you will predict. The unit is USD. This column doesn't exist in test.tsv since that is what you will predict.
- shipping - 1 if shipping fee is paid by seller and 0 by buyer
- item_description - the full description of the item. Note that we have cleaned the data to remove text that look like prices (e.g. $20) to avoid leakage. These removed prices are represented as [rm]

Using the given data, the goal is to predict the sale price of a listing based on information (features) that a user provides for this listing.

The approach of this project will follow the standard data science project life cycle. First, start with understanding the problem. Then, data wrangling (looking for missing values, incomplete data, etc.) follows. After that, exploratory data analysis will follow to get more understanding of the data and to identify patterns, outliers, etc. For this particular project, the exploratory data analysis will include text analysis (item description column). Then, fit the training data to one or more predictive machine learning models to perform the price prediction. After that, perform parameter tuning and optimize the model for the best accuracy. The deliverable of the project will be a code containing all the analysis and the algorithm and a complete project report explaining the problem, the goal, the dataset, the model, the steps taken and decisions that are made to create the model, and the results.