Project Proposal - Team 27

Arnav Saxena (as6456)
Nuanyu Shou (ns3492)
Oliver (Hongiu) Liu (hl3518)
Xinfu Su (xs2444)
Zihao Liu (zl2986)

Background & Context

Mercari is a Japanese e-commerce company. Their main product is the community-powered marketplace they offer via the Mercari marketplace app. It's just like eBay where people can buy and sell used products ranging from clothing to electronics and more.

The main objective of our proposed project will be to use data of items sold in the past and build a model that could suggest the right product prices to sellers.

Datasets

Mercari hosted a kaggle contest around the same topic a few years back. We plan to use the data they provided for the same.

Link: https://www.kaggle.com/c/mercari-price-suggestion-challenge/data

The data provides the following information for about 1.5 million SKUs:

1. **train_id, test_id** — the id of the product
2. **name** — the title of the product
3. **item_condition_id** — the condition of the product provided by the sellers
4. **category_name** — category of the product
5. **brand_name** — the product's brand name
6. **shipping** — 1 if shipping fee is paid by seller and 0 if shipping fee is paid by buyer
7. **item_description** — the full description of the product
8. **price** — the price that the product was sold for

Proposed ML techniques

We are dealing with what clearly is a **supervised regression problem**. We aim to start with traditional regression models such as ***linear regression, RandomForest, LGB and Xgboost***, and eventually progress towards more nuanced ***deep learning models***.

Furthermore, given that we have a very important free text variable available in the data (*“item_description”),* we might try to extract informative features from text by using techniques such as **TF-IDF** and **word embeddings** (word2vec/FastText**). Transformer** may also be a possible framework to use.

In a nutshell, we plan to utilise deep learning in the following two ways while solving this problem. Firstly, by extracting high level features from the original categorical and textual columns. And secondly, by applying neural networks for direct end-to-end training.

Besides, we will try to combine traditional and deep models for final prediction.