Capstone proposal

Domain Background

After extensive research, I propose to run a Kaggle completion on Mercari Price Suggestion Challenge as my capstone project (https://www.kaggle.com/c/mercari-price-suggestion-challenge).

Mercari (http://www.mercari.com) is Japan's biggest community-powered shopping application, similar as eBay in United States. They would like to offer pricing suggestions to sellers. In this competition, Mercari is challenging us to build an algorithm that automatically suggests the right product prices. I will predict the sale price of a listing based on information a user provided for this listing.

Problem Statement

It can be hard to know how much something is really worth online. Small details can mean big differences in pricing. In addition, the sellers may just put about anything or any bundle of things, so that it is tough to suggest the price to sell.

Datasets and Inputs

I will be provided with user-inputted text descriptions of their products, including details like product category name, brand name and item condition.

Train and test files in .csv format are provided. These files are tab-delimited. Both files are not too large so that the data can be accessible for my reviewer.

Training file is 342.4MB with 8 columns and has 1,482,535 records, and the test file is 170.9MB with 693,359 rows.

Solution Statement

As the price prediction is continuous value, this is a supervised machine-learning problem.

In this case I will run the following models to compare which model can generate the best performance:

1. Keras Neural Network
2. XGBoost with linear booster,
3. Random Forest model
4. CatBoost model.

Benchmark Model

As this is a supervised problem, I will run a linear regression model as the benchmark model so that I can check how the advanced complicated models performance.

Evaluation Metrics

The evaluation metric for this competition is Root Mean Squared Logarithmic Error.
The RMSLE is calculated as below:

$$e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(pi\ +1) - \log(ai+1))**2}$$

Where:
e is the RMSLE value (score)
n is the total number of observations in the (public/private) data set,
pi is your prediction of price, and
ai is the actual sale price for i.
log(x) is the natural logarithm of xx

Project Design

As the datasets will be provided by Kaggle, I can download the dataset and run the algorithms on my local computer. My computer already have tensorflow, xgboost and catboost models installed, so that no additional software needed.

I will run the whole project on Jupyter notebook on python 3.5.
My report will include the following sections:

1. Explanatory Data Analysis
    - Load the data and fill the missing values
    - Check training data description
    - Check the target variable: price distribution
    - Check number of items by main category and subcategory
    - Price distribution by general category
2. Text processing on Item Description
    - Tokenization for item description
    - Visualization which words has the highest frequency
    - KMeans clustering of the item description
3. Base Model as benchmark
    - Linear Regression Model
    - Compute RMSLE
4. Advanced Models
    - XGBboost
    - Keras Nerual Networks
    - Catboost
    - Random Forest Model

Project Summary

In this section, I will summarize what I skill I polished and what lessons I learned while complete this capstone project.