Team:

Durga Chillakuru
Dipanjan Haldar
Chandrashekar Gubbi MurugendraPrasad.

Problem Description:

Pricing Strategy: The challenge of setting the right price for a product

Setting the right price for a good or service is an old problem in economic theory. One company may seek to maximize profitability on each unit sold or on the overall market share, while another company needs to access a new market or to protect an existing one. Moreover, different scenarios can coexist in the same company for different goods or customer segments.
We are trying to present the problem of **price optimization for retail** and how retailers can take advantage of the tremendous power of Machine Learning (ML) technology to build effective pricing strategy **using dynamic pricing model**.

What are the inputs and outputs (exactly)?

Input: We plan to use a dataset used to solve the dynamic pricing challenge. We plan to use the Mercari dataset that was given as a part of the Mercari-price-suggestion-challenge at Kaggle. The data consists of descriptions of the products, including details like product category name, brand name, and item condition.

Output: Build an algorithm that automatically suggests the right product prices. We plan to use 3 different algorithms to predict the prices and will finally compare them to select the best algorithm for price prediction. We plan to use RMSE (root mean square error) which is a measure of accuracy to compare the results from different models for the given dataset

What data are you using (exactly)?

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

Algorithms:

- KNN, KMean (Durga)
- Linear Regression, XGBoost (Dipanjan)
- Light GBM, Ridge Regression (Chandra)

Why are these algorithms appropriate? How are these algorithms typically used, and how are you using them?

K Nearest Neighbors - **KNN** is a **non-parametric, lazy** learning algorithm. By non-parametric, it means that it does not make any assumptions on the underlying data. So, basically KNN is the first choice for a classification study when there is little knowledge of the data. A positive integer k is specified, along with a new sample. We select the k entries in our database which are closest to the new sample. We find the most common classification of these entries. This is the classification we give to the new sample.
https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/

K Means – K-Means is one of the most popular "clustering" algorithms. K-means stores kk centroids that it uses to define clusters. A point is in a particular cluster if it is closer to that cluster's centroid than any other centroid. A cluster refers to a collection of data points aggregated together because of certain similarities. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
The word *'means'* in the K-means refers to averaging of the data; that is, finding the centroid. In our dataset we would be taking the categories as centroids and try to classify the data points which are closest to one of the centroids. This algorithm will help us in to find segmented pricing

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Linear Regression - It is a type of regression analysis where there is a linear relationship between the independent(x) and dependent(y) variable. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). In our case, we plan to use this algorithm to determine the relationship between Price (independent variable) and other dependent variables like category, brand, shipping, item condition etc. and how these factors have an impact on the independent variable, price.
https://machinelearningmastery.com/linear-regression-for-machine-learning/

Light GBM - Light GBM is a gradient boosting framework that uses tree-based learning algorithm. Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithms grows level-wise. It is best suited for large datasets and it is highly efficient in terms of speed and consumes lower memory to run.
https://towardsdatascience.com/a-case-for-lightgbm-2d05a53c589c

Ridge Regression – Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares

estimates are unbiased, but their variances are large, so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.
https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf

XGBoost - It is based on an end-to-end tree boosting system. It offers a faster and more accurate way to solve classification and regression-type problems. The key feature in XGBoost is that it weights the predictors and tries to keep the new decision tree away from the errors made by previous decision trees. Thus, it strengthens its accuracy. During the review of other professional research papers, we found that XGBoost is well adapted for dynamic pricing problems, which is relative to our purpose - predict the price for online retailers.

https://mwdsi2018.exordo.com/files/papers/84/final_draft/XGBoost_-_A_Competitive_Approach_for_Online_Price_Prediction.pdf

https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Have other people use similar algorithms to solve your problem before?

Yes, people have used these algorithms for predicting the price.

Results:

What results do you expect to show?

In our project we are planning to use td-idf vectorizer to predict the weightages of each feature in the given dataset. This weightage will help in identifying the required feature from the given data set. The obtained sparse matrices will be merged with CSR matrix. CSR matrix provide better memory utilization and efficient row slicing.

We will generate the Root Mean Squared Error for different algorithm by varying the number of features we selected during tf-idf vectorization and propose an algorithm which best describe the solution for dynamic pricing problem.

What comparisons will you do?

We will perform a comparison between the various models (Linear Regression, Ridge regression, Light GBM). Based on the RMSE's generated for various models, the best model will be selected to suggest prices

Are there risks for not getting all the results? If so, what will you do about it?

We are planning to use KNN classification or No classification to clean the data sets. Since the data set is too large, we might encounter some risk in cleaning and preprocessing. This might force us to use limited data set for the prediction.

References

https://tryolabs.com/blog/price-optimization-machine-learning/
https://medium.com/@RemiStudios/artificial-intelligence-for-dynamic-pricing-326c3b88a37
https://datascienceplus.com/knn-classifier-to-predict-price-of-stock/
https://www.academia.edu/38450105/Learning_market_prices_in_real-time_supply_chain_management
https://www.kaggle.com/c/mercari-price-suggestion-challenge