Market Trend Analysis and Prediction Model

- Using ML and Statistics

ABSTRACT

Datasets and Corpora have been systematically collected, sorted and combined to form meaningful information sets. These sets, built using the right models, provide an insight into a user's searching, buying and spending patterns. This allows us to analyse and predict the future activities of a user, and on a larger scale - an Industry.
Change being a constant factor in the fashion industry, it can be hard to predict changes in price, inflation, designs, patterns, etc. However, with the use of Machine Learning, we can train models to not only analyze, but also predict these changes over a reasonable margin of error. This helps in efficient and user-friendly pricing for an industry, and gives a user insight into the expected pricing for a commodity.

This Project aims to build regression model over sample datasets to get an insight into the fashion industry. Regressions help us understand the trend of an industry, analyze losses and factors contributing to them. Clustering a group of similar traits helps in building efficient pricing models, which is what is aimed for the project.

Keywords

Machine Learning, Statistics, Regression, Dataset, Model

Introduction

In the recent years, due to cheap and easily accessible internet , e-commerce industry has flourished. Online industries are no longer limited to regular shopping stops - they sell anything and everything while providing a host of services, reducing time needed to sort through each item in a physical shop. However, better e-commerce experience can be provided if user's search and buying patterns are monitored and analysed. This rising age of information trade has given way to formation and maintenance of data-sets, provided one can sort and take valuable inferences. This paves way for market research for trend analysis, for better user experience and faster , more efficient item pricing. We can use Machine Learning and Statistical Analysis to make models for prediction and trend analysis.

Need for Trend Analysis

Markets for user-consumption regular wear and even fashionable clothing are very fickle. Trends tend to change as time progresses. Many a times, an older trend becomes more fashionable, or new fashion insights are taken from a current trend. Change is the only constant in this industry, and as Markets change - people's interests tend to shift in their favor.
The Fashion and Clothing industries are heavily celebrity influenced, even for regular comfort wear. The populous looks up to idols and wants to adorn the way of life that they do - albeit on a meagre scale.

However, Trends change as market shifts in favor of celebrities all over the world. Sometimes influenced by the fashion towards the eastern side of the world, sometimes from the west. Fashion does not stay constant for a long period of time, and hence trends shifts are quite common. Since each trend does not cater to the same market value and pricing, a change in trends always accompanies a change in price for clothing. A seasonal change is to be expected at each of the following trends. Being able to keep up with the market on a very minute scale might not be the most easy solution to stay with the times - however, this ambiguity in change can be predicted using Machine Learning. An ML model can fit, predict and analyse with data inferences all quantifiable measures.

Working

Datasets for previous sales over platforms like Flipkart, Amazon, Myntra etc. provide an excellent resource for previous trends. An ML model based on these trends can predict a price-value variation on unseen data, which will provide the basis for our learning algorithm.

Cleaning of the data-set to extract needed features is required to start working with any given dataset. An ML model using the following can analyze and predict data as we train it :

i. Linear Regression
ii. Logistic Regression
iii. SVRs (Support Vector Machine with regression)

SVM clustering removes chances of overfitting data, so there is no need for regularization. However, due to inflation in markets, normalization and scaling of prices may be required to get required accuracy for the model.

Data inferences for popular items and user-age inferences let us know what age group prefers a certain item, so we can further cluster w.r.t. Age metrics for better UX.

However, a better idea for dummy model data is a combination of clustering and providing a regression for it. A Random Forest approach or a Support Vector Regression is able to classify and give a steady, accurate reading. We have used SVR as a combinational algorithm for predictions.

Future work for the model is aimed at using targeted Sentiment Analysis with Tanimoto Similarity and WESM ranking systems, to decide against 2 equally preferred trend models. Including news reports for fashion trends all over the world, GDP pricing for country-wise inflation to decide the best country to buy an item with the cheapest price but same quality, etc is aimed to make the model a filtered tight-bound, so that it can predict prices with minimal accuracy loss.

Two Modules, linmodel.py and Trend.py are used to make analysis and predictions respectively. Combined, these two form a trained model.

Metadata Information

The metadata collected can be divided into two sets for working , **Analysis** and **Prediction**.

For **Analysis** , the dataset collected is cleaned and worked with using Pandas. The datasets are compiled from kaggle to train and test the data, Myntra to try real world examples and

Flipkart / Amazon to make future predictions on country - scaled values. A self-scraped dataset with minimal feature selection was also made to simplify the process.
A dummy dataset was used to prepare a prediction model , since real-time Prediction model datasets were unable and combining datasets was inefficient due to insufficient online records. A stock prediction model was taken as a reference and random, dummy data was entered for a sample model.
The **Prediction** model dataset is cleaned using the Backward Elimination model building technique , used as a Feature Selection technique to maximize prediction accuracy while maintaining an average computational complexity. The sample data was cleaned with Pandas library. Using multiple libraries, encoding and scaling of features was performed to plot data inference graphs. Weighted Multilinear Regression and SVRs were used independently to provide predictions.

Related Work

Machine Learning is mainly used for Regression Plotting, Classification and Market Inferences in Data Analytics. We have combined and cherry-picked techniques from the above categories to make a stable, working real-time model. The proposed model uses the regression techniques mentioned in the *Working* subsection, and each of them can be explained as shown below :

Linear Regression is a way to model the relationship between two variables. You might also recognize the equation as the **slope formula**. The equation has the form Y=a+bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

The first step in finding a linear regression equation is to determine if there is a relationship between the two variables. This is often a judgment call for the researcher. You'll also need a list of your data in x-y format (i.e. two columns of data — independent and dependent variables).

Multiple linear regression analysis predicts trends and future values. The multiple linear regression analysis can be used to get point estimates. An example question may be "what will the price of gold be 6 month from now?"
When selecting the model for the multiple linear regression analysis, another important consideration is the model fit. Adding independent variables to a multiple linear regression model will always increase the amount of explained variance in the dependent variable (typically expressed as R^2). Therefore, adding too many independent variables without any theoretical justification may result in an over-fit model.

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

where R is the coefficient of determination. SSres and SStot are sum of squares of residuals and total sum of squares.

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2 \qquad SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

Support Vector Regression can be subdivided into two types : Linear SVR and Non-Linear SVR. Due to the randomness of the dataset used, we have used Non-Linear SVR for model fitting. Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences.
For Non-Linear SVR, the kernel functions transform the data into a higher dimensional feature space to make it possible to perform the linear separation. The SVR, mathematically, can be explained as :

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot \langle \varphi(x_i), \varphi(x) \rangle + b$$

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot K(x_i, x) + b$$

Visualisation

To better understand the data provided by the dataset, visualisation is done by plotting various graphs on the parameters that are present in the dataset. Multiple graphs were plotted with respect to inflation, seasonal pricing, discounted pricing and year stated. Sample Data Plotting was done for both, Analysis and Prediction models.
The Prediction Dataset used is a dummy set of 1000 observations of inflation, seasonal price changes and discounted pricing. The Analysis set is real-time data for accurate pricing distribution. This yields a linear correlation, which can be seen in the *Results* section.

Results

The Analysis Dataset has data-fields spanning 15 categories. Out of these, meaningful inferences for analysis were drawn from only 2 fields. As a result, a graph plotted on Analysis model of the data contains a visualization of only these fields.

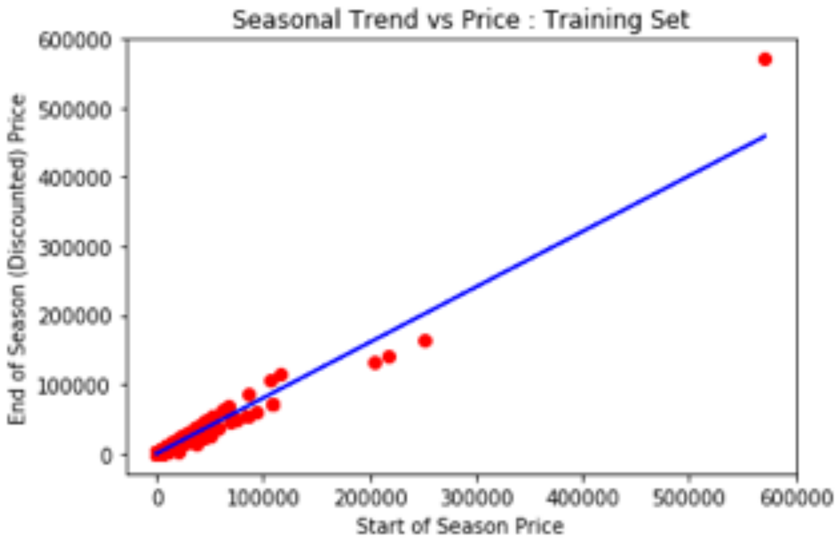

Fig 1.1 : Training Set Data Graph for Model - 1

As we can infer, the scatter plot shows a direct correlation in seasonal start and discounted prices at season end. The same regression line was used to plot against the testing set, the results of which can be seen below

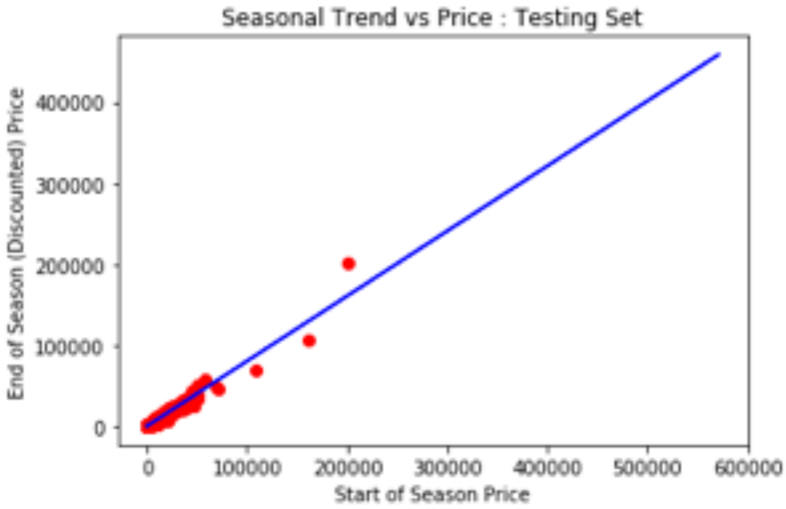

Fig 1.2 : Testing Set Data Graph for Model -1
The visualization of inferences for Model 2 is as shown below:

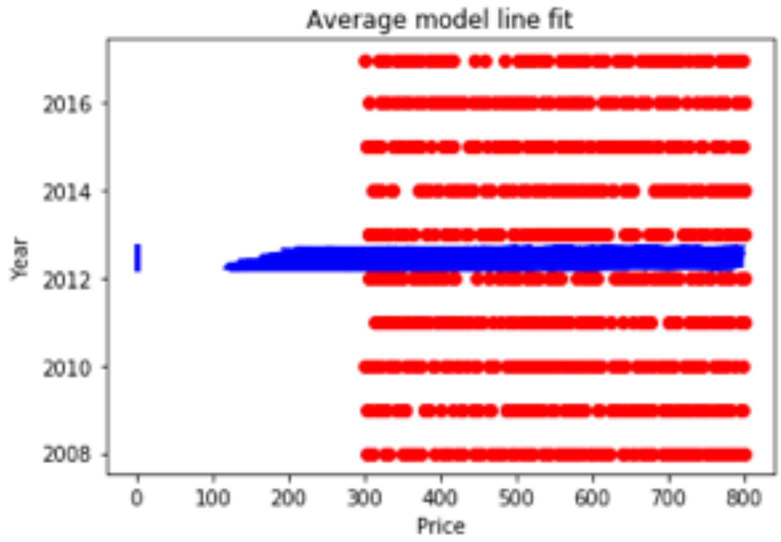

Fig 2.1 : Model line fit using Multiple Linear Regression over Model - 2

As the sample data given was a dummy set, the model fits the most consistent loss model based on coefficient of determination. A sample regression fitting can be seen below :

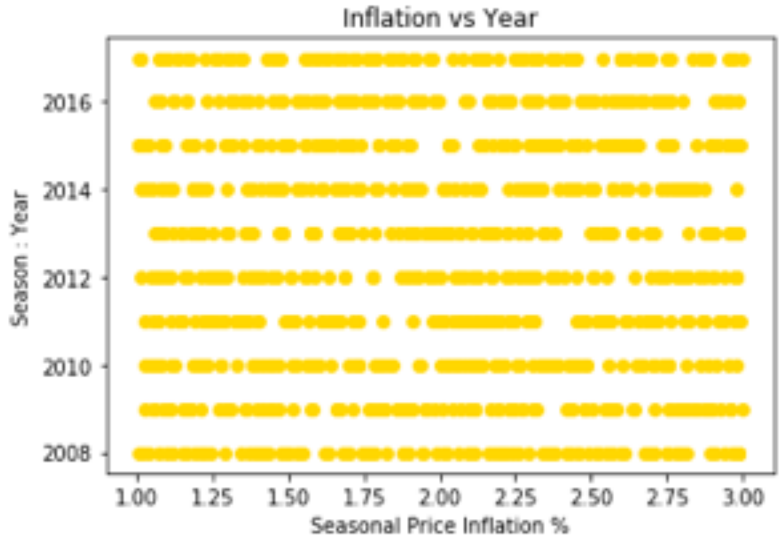

Fig 2.2 : Random Inflation setting over the decade 2008 - 2018 pertaining to Model - 2

The SVR model gave a mode strong-held value for the prediction value, covering less data-points in consideration with loss and coefficient of determination (R^2). A strong-held value comprising of SVR and MLR models can be cross-verified with model fitting. The SVR graph over sample data is as shown below:

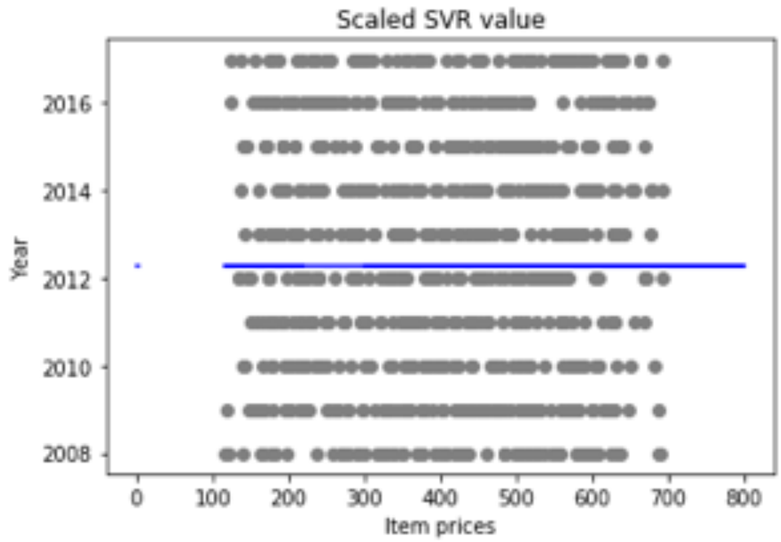

Fig 2.3 : SVR (scaled) over Dataset for Model - 2

This Market Trend Analysis and Prediction model can be used for different consumer related services, such as Software building, Finances, E-commerce, etc. This an exploratory form of Data Analytics, and can be researched into further. The given model is only a sample idea for the project, however, a more complex system can be built using multiple metrics, provided a reliable data-set is procured.

Libraries Used :
Pandas, Numpy, Scipy, Scikit-Learn, Matplotlib. All libraries are in Python

Framework :
(For WebDev) Flask

Languages Used:
Python (ML model)
HTML, CSS, Javascript, Flask(Webdev)