Mobiles Sales Strategizer

Srish Srinivasan
Computer Science Engineering
PES University

Bangalore,India
srishwap4@gmail.com

Sriram Subramanian
Computer Science Engineering
PES University

Bangalore,India
sriramsub7@gmail.com

Abstract

This paper discusses the various methods that are employed in predicting the trends in sales data. Here, an attempt to forecasting mobile phone sales has been made. Initially, we learned how Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) have been used to predict the demand for electricity in a residential locality. We then saw how the application of Deep Neural Networks in building predictive models came up with some very accurate predictions. Moving on, we also looked at how Support Vector Machines have been employed in forecasting the sales of mobile phones. We then go on to implement models such as simple linear regression, multiple linear regression, random forest regression and ARIMA to analyze the sales data and also try and find out the importance of features.

I. Introduction

A market place is a intermediary link between a merchant(s) and a consumer(s). It takes away the hassle of visiting different websites or visiting different malls in order to compare the pricing of a product that a consumer is interested in. So a market place, be it online or offline acts as a portal for all merchants to exhibit and sell their products to interested customers.

Customers would find it really easy to make a sound decision on their purchases given the plethora of options in terms of pricing and features. The merchants too could analyze and find out the products that seem to be really trendy and accordingly plan their sales strategy and the investments that they would have to make in order to have a successful campaign at the marketplace.

Our goal was , given the data set of phone sales in the first half of 2020, try and analyze how certain sales decisions such as pricing were indeed getting the optimal results in terms of sales, try predicting sales of a product and try to bring out the most influential factors that affect the sales.

The data was considerably clean when we ingested it, therefore not much of cleaning was required.

One of the issues was, that many of the columns in the data set took unstructured values, and we had to figure out a way to extract information from them or make the decision to drop the columns entirely.

Exploratory analysis showed that many attributes had a skewed distribution. Few of the attributes were considerably correlated with one another. For one, the *daily_gmv* attribute of the data set was highly correlated with the *daily_sold* attribute.

II. Literature Review

Anupiya Nugaliyadde, Upeka Somaratne, and Kok Wai Wong Murdoch in their paper “Predicting Electricity Consumption using Deep Recurrent Neural Networks” talk us through their approach to predicting Electricity Consumption using a Recurrent Neural Network (RNN) and a Long Short Term Memory network (LSTM). Both these models are purely dependent on the historical consumption of eletricity. They observed that the analysis of time-series data through conventional approaches like ARIMA (Autoregressive integrated moving average), Fuzzy based techniques, SVM (Support Vector Machines) and a few others did an excellent job in making short-term predictions but were considerably poor when it came to making long-term predictions. RNN and LSTM based models consist of a feedback loop from their past inputs in order to learn from the previous sequences of data. Both these models were observed to minimize the RMS (Root Mean Square) error value to 0.1 for all the cases on an average. They concluded that the RNN and LSTM based models performed as good as ARIMA in making short-term predictions and outperformed all the other models when it came to making medium-term and long-term predictions with very good accuracy. **[1]**

In the paper , “deep neural decision trees”, Yongxin Yang, Garcia Morillo, and Timithy M.H , discuss how neural networks are extremely powerful, but lack the interpretability of tree based models. In use cases corresponding to fields such as business and medicine, the model may require to be explainable. They incorporated both the decision tree and ANN techniques to create a custom model. This model, instead of using greedy techniques to grow the tree, uses stochastic gradient descent and can hence find the better approximate to the optimal solution. Since the entire model is constructed like a decision tree, It would be quite easy to interpret. They tested on tabular, low dimensional data and compared the performances of a dndt, decision tree and a neural network. They concluded that the decision tree slightly outperforms the dndt as it is more suited for the type of data used, but the dndt did outperform the neural network . Maybe the dndt could provide an useful, interpret able model when analyzing perceptual data. **[2]**

Zekun Duan in his paper “Mobile Phone Sales Forecast Based on Support Vector Machine” discusses how he employed SVM (Support Vector Machine) in building a model for the prediction of mobile phone sales. He took in factors such as price, wear resistance, resistance to fall, charging interval, battery life, communication stability, photo effect, appearance design, memory size and the willingness to buy again as input variables and the sales foreground grade as the output variable. The reason behind employing SVM was that it was very efficient in making predictions as it overcomes the problem of local optimum and also eliminates the possibility of a long convergence time which seems to be very evident in some of the neural networks based models. On analyzing the predictions that were made, he concluded that the sales forecasts were very consistent with the actual values announced by the mobile phone manufacturers. **[3]**

III. Problem Statement

1. To find out the most influential independent variables in determining the value of the dependant variable.

2. To build a model that can predict the sales of a mobile phone given its parameters.

3. To estimate the amount of history to consider when making a prediction , since it is time series data.

IV. Approach Overview

Since the data set contains records of numerous products, to provide valuable insights analysis was conducted on subsets of the dataset corresponding a particular product type. The product type to be selected should be configurable and hence analysis can be done all products, one at a time.

To achieve preliminary insights, various plots would be illustrated.

We would be using regression models as the target variable is a continuous variable.

Since the target variable varies with time, methods for time-series analysis could be employed.

The models of simple linear regression, multiple linear regression, random forest regression and auto regressive integrated moving average were used for analysis.

The importance of the various features involved were extracted.

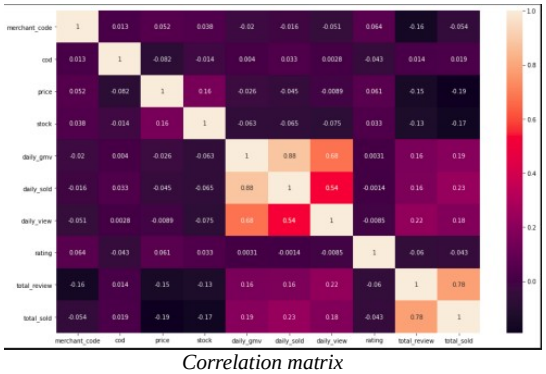
The best performing model was chosen and k-fold cross validation was conducted.

V. Our Approach

All further analysis was done only on a particular subset of the entire data set. This subset corresponds to all data of a particular product type. This was accomplished using a function which will take in the product type as a parameter and hence the user could experiment and analyze the various product types. This function returns a processed subset.

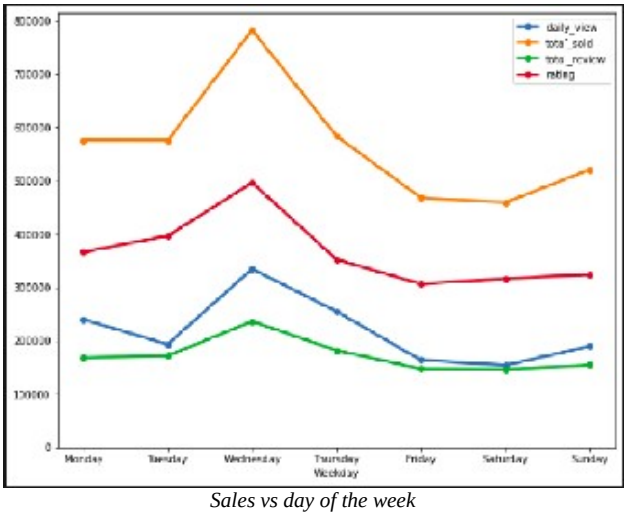
1. Exploratory Analysis

The data had very few missing values. Hence, records with missing values were discarded. Exploratory analysis was conducted before pre-processing to get a better idea of the distribution of data. As well as to judge the outliers in the data and bring out any interesting characteristics.

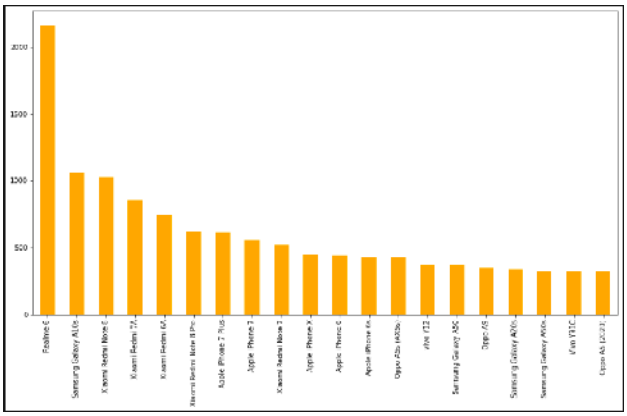


Correlation matrix

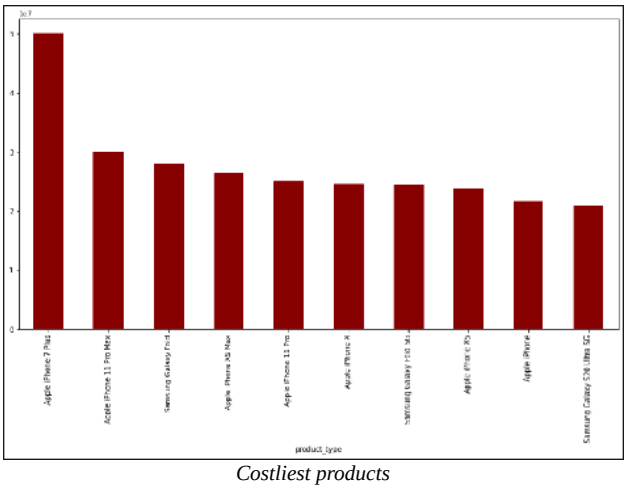
Box plots were plotted for each feature. The correlation matrix was plotted to see how the features were related to each other. Various parameters of the data were grouped by the days of the week and plotted. Plots such as the top 20 brands with the most number of products, the 10 costliest products and the total products sold against the dates were illustrated.



Sales vs day of the week



Brands with most number of products



Costliest products

2. Pre-processing the data

Missing values were already taken care of before the plots were made. Further processing was done to ready the dataset for training. The range of the different parameters varied greatly, so the data had to be scaled before further processing. The categorical variables had to be encoded. Features that were not relevant to the analysis were discarded. Records with outliers in any of the features were discarded.

testing

3. Building the models

A processed subset of data corresponding to a particular product was retrieved.

The dependant variable was identified to be the total_sold feature.

The feature that was the most correlated with dependant variable was total_review. This is understandable as only the customers who buy products would review them. This must be considered during any further analysis.

i. Simple Linear Regression

A simple linear regression model was trained with total_sold as the dependant variable and total_review as the independent variable. This was done to further judge the relationship between the 2 variables. It yielded an R2 score of 0.98244 with an mean absolute error of 11.15 and a mean squared error of 247.22 on the test dataset. This shows that total_review plays a major role in determining the value of total_sold.

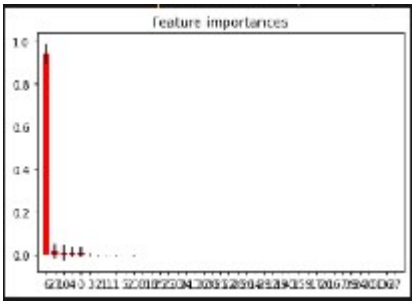
Training

ii. Multiple Linear Regression

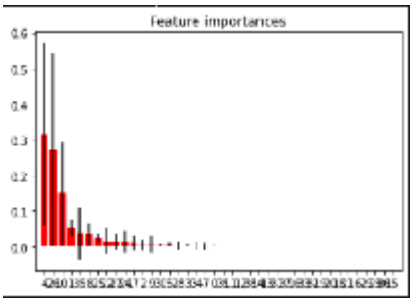
A multiple linear regression model was trained with total_sold as the dependant variable and the independent variables being price, stock, daily_gmv, daily_sold, daily_view, rating, total_review and the encoded dummies of cod and merchant_city. It yielded an R2 score of .997313 and a mean absolute error of 5.04 and a mean squared error of 42.76. However, since total_review should not be considered for analysis, the model was trained again during k-fold cross validation without it being considered as an independent variable.

iii. Random Forest Regressor

A random forest regressor model was trained with total_sold as the dependant variable and the independent variables being price, stock, daily_gmv, daily_sold, daily_view, rating, total_review and the encoded dummies of cod and merchant_city. It yielded an R2 score of .9745 and a mean absolute error 12.46 and a mean squared error of 406.39. The trained model was also used to extract the importance of the features according to the model. As expected, total_review was the most important feature. Hence, the model was trained again without considering total_review as one of the dependent variables. This model yielded an R2 score of .83287 and the importance of features was extracted again and daily_view turned out to be the most important feature. The product_id played a major role in determining the dependant variable . The features stock and daily_gmv also ranked high in importance.



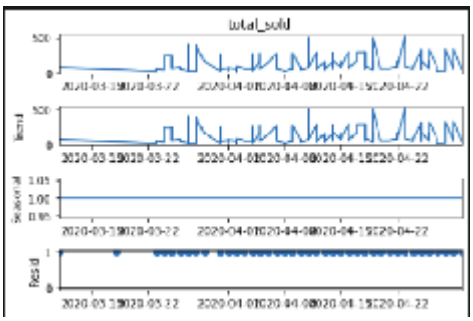
With total_view feature



Without total_view feature

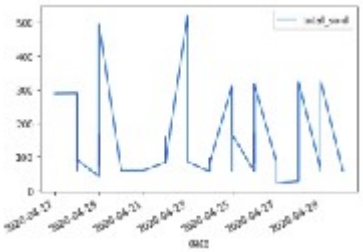
iv. ARIMA

We decomposed the signals in the variable total_sold into trend, seasonality and residual components. No significant trend or seasonality was observed. Hence, it was suitable for the application of arima.

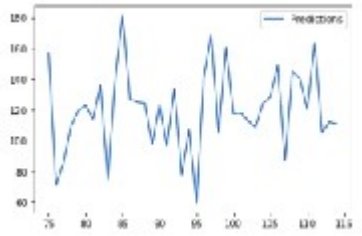


Decomposition of signal

The appropriate values for the parameters p, d and q were extracted from the data. The model resulted in a root mean squared value of 134.21 and an mean squared value of 18013.19. This clearly showed that the model was a complete misfit for the given data.



Observed test data



Predictions by the arima model

v. Feature Importance

To confirm the results given by the random forest model, the highest ranked features using the f_regression score were extracted. The rankings were same as that output by the random forest regressor.

vi. K-Fold Cross Validation

The model with the best performance turned out to be Multiple Linear Regression. This model was validated by considering with and without the feature total_review during training. The model that considered total_review as one of the independant variables yielded R2 scores in the range .9960 to .9985 on 7 splits. The model that hadn't considered total_review as one of the independent variables yielded R2 scores in the range .7413 to .9867 on 7 splits.

VI. Insights

1. The multiple linear regression model was observed to be the most effective model among the ones that were tested.

2. The daily views of a commodity strongly determines what the future sales will trend like.

3. Stock and price play a major role in determining the sales as well.

V. Conclusions

Contributions made by SRN PES1201800051, Srish Srinivasan:

Explorative analysis and plotting, simple linear regressison, auto regressive intgerated moving avearge method.

Contributions made by SRN PES1201800655, Sriram Subramanian:

Multiple linear regression, Random forests, feature extraction and k-fold cross validation.

Note

Conducting a project with 2 members had its pros and cons. Coordination was definitely very easy. But it would have helped to have more personnel on the fronts of research and implementation.

References

[1] “Predicting Electricity Consumption using Deep Recurrent Neural Networks” by Anupiya Nugaliyadde, Upeka Somaratne, and Kok Wai Wong Murdoch

[2] “deep neural decision trees” by Yongxin Yang, Garcia Morillo, and Timithy M.H

[3] “Mobile Phone Sales Forecast Based on Support Vector Machine” by Zekun Duan