PREDICT FUTURE SALE

Ramya Nagalla	**rnagalla@kent.edu**	**811122347**
Likitha Chowdary Pasam	**lpasam@kent.edu**	**811132247**
Prashanth Reddy Challa	**pchalla1@kent.edu**	**811128902**
Varun Reddy Pisati	**vpisati@kent.edu**	**811169288**

ABSTRACT

For modern retail businesses with a large chain of stores, accurate sales forecasting is criticalto the company's growth, as well as its success or failure. Sales forecasting enables businessesto better allocate resources, such as cash flow and production, and to make more informed business decisions. The management of stores and products, or "supply chain management," is an overly complex process. It is not practicable to operate stores without knowing what inventory is required to manage and run the store smoothly. One of the most critical aspects in keeping the store open is predicting sales. Sales forecasting is critical for store management and profit. It is an important aspect of business administration. The research focuses on estimating future sales for the next immediate month utilizing time series previous sales data, assisting businesses in inventory management. A Future sales prediction dataset provided by Russian software firm – 1C collected from Kaggle is being used.

Introduction:

Predicting sales is one the most important steps to keep the store going. Forecasting sales is critical for managing the store and increasing revenues. It is a necessary component of company management. You can't manage your inventory, cash flow, or plan for expansion until you know what your future sales will be. The goal of sales forecasting is to give information that may be used to make informed business decisions. We are currently living in a period of dramatic transformations driven by digitalization, information and communications technology, artificial intelligence, and other technologies. Many business and economic academics believe that this will herald in a new age known as the Fourth Industrial Revolution. The decision-making process will be the most important shift in this social upheaval.

This is evident in the retail industry, where modern technologies allow businesses to forecast sales accurately. As a result, businesses are better able to make sound decisions and maximize their resources. Take Walmart as an example: by applying machine learning for sales forecasting from vast historical data, they can improve their cash flow, staffing, production,and financial management. It can also help investors by reducing uncertainty and anticipating market movement.

Project Techniques for creating future projections based on current and historical data has always been a field with direct application to a variety of real-world issues. We're talking about a comparable issue right now. In this project, Machine Learning techniques were used to aid with a similar scenario faced by a huge Russian software corporation, 1C, which is attemptingto anticipate total sales for every product and store in the coming month. The company has given you a difficult time series dataset with daily sales data. This project focuses on predicting sales using several methods.

The main goal is to forecast total sales for each product and store in the coming month, as well as product sales in relation to retailers for the test set. As a result, a robust and dynamic model that can withstand dynamically changing time series sales data and reliably anticipate future product sales has been developed.

Workflow

1. Project Life Cycle

In this project we have the following phases Data Collection, Exploratory Data Analysis (EDA), Data Pre-processing & cleaning, Model building, Model Evaluation. Below diagram illustrates the project life cycle.

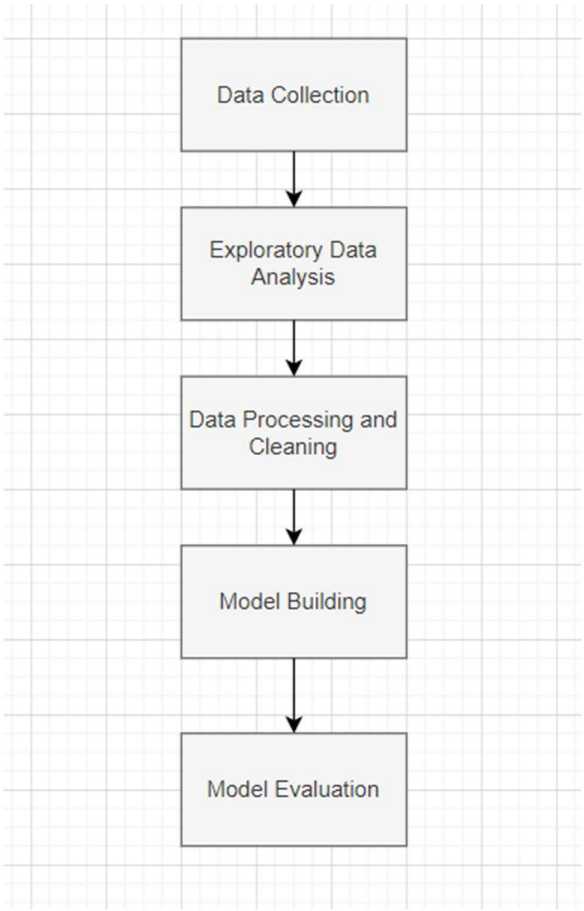


Figure 1 Life Cycle

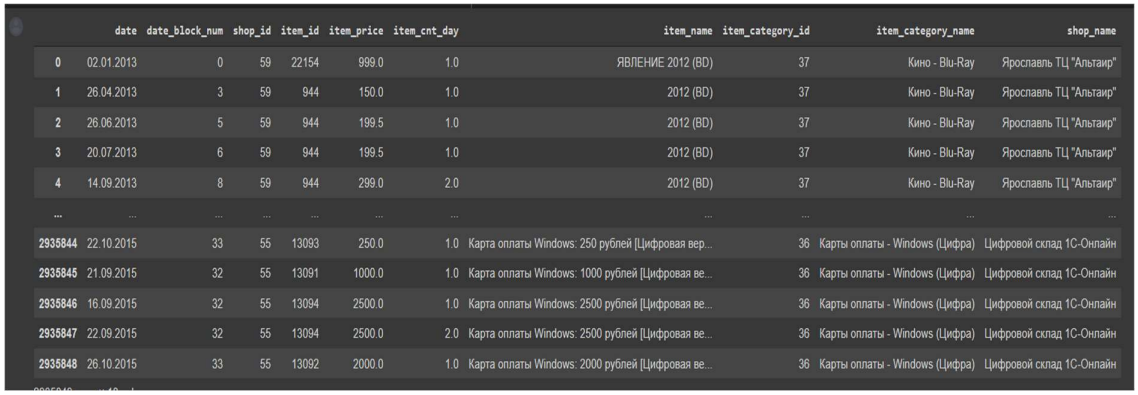
Data Collection

The dataset is a time series data of the daily sales of products in multiple stores, which is provided by the Russian firm -1c. The data is obtained from kaggle. The dataset that

was used contains daily historical sales data of over 22,000 items from 60 different shops for a date range from January 2013 to October 2015.

URL : https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data

Snapshot of raw data before preprocessing:



	date	date_block_num	shop_id	item_id	item_price	item_cnt_day	item_name	item_category_id	item_category_name	shop_name
0	02.01.2013	0	59	22154	999.0	1.0	ЯВЛЕНИЕ 2012 (BD)	37	Кино - Blu-Ray	Ярославль ТЦ "Альтаир"
1	26.04.2013	3	59	944	150.0	1.0	2012 (BD)	37	Кино - Blu-Ray	Ярославль ТЦ "Альтаир"
2	26.06.2013	5	59	944	199.5	1.0	2012 (BD)	37	Кино - Blu-Ray	Ярославль ТЦ "Альтаир"
3	20.07.2013	6	59	944	199.5	1.0	2012 (BD)	37	Кино - Blu-Ray	Ярославль ТЦ "Альтаир"
4	14.09.2013	8	59	944	299.0	2.0	2012 (BD)	37	Кино - Blu-Ray	Ярославль ТЦ "Альтаир"
...	...	...	...	...	...	...	...	...	...	...
2935844	22.10.2015	33	55	13093	250.0	1.0	Карта оплаты Windows: 250 рублей [Цифровая вер...	36	Карты оплаты - Windows (Цифра)	Цифровой склад 1С-Онлайн
2935845	21.09.2015	32	55	13091	1000.0	1.0	Карта оплаты Windows: 1000 рублей [Цифровая ве...	36	Карты оплаты - Windows (Цифра)	Цифровой склад 1С-Онлайн
2935846	16.09.2015	32	55	13094	2500.0	1.0	Карта оплаты Windows: 2500 рублей [Цифровая ве...	36	Карты оплаты - Windows (Цифра)	Цифровой склад 1С-Онлайн
2935847	22.09.2015	32	55	13094	2500.0	2.0	Карта оплаты Windows: 2500 рублей [Цифровая ве...	36	Карты оплаты - Windows (Цифра)	Цифровой склад 1С-Онлайн
2935848	26.10.2015	33	55	13092	2000.0	1.0	Карта оплаты Windows: 2000 рублей [Цифровая ве...	36	Карты оплаты - Windows (Цифра)	Цифровой склад 1С-Онлайн

Data Preparation

In Data Preparation process we do the following steps.

- Missing Values Evaluation
- Removing Outliers

Missing Values

The first step in this is to detect the missing values and to correct them from the given data set. Then the numerical values are treated with Median. Then the categorical missing values are treated with the “Missing” string.

Missing Values Evaluation

The dataset collected from Kaggle was cleaned, and it was discovered that there were no missing or null values in the dataset.

```
train_df.isnull().sum()

date              0
date_block_num    0
shop_id           0
item_id           0
item_price        0
item_cnt_day      0
dtype: int64
```

Figure 2 – Data Set with no missing values

Outliers- This process performs all necessary data preprocessing and model optimization. Outlier detection process can be used to deploy the model or as a starting point for further optimizations and helpful in showing generic information which is independent of the models. The focus is on the quality of the data, especially the quality of each data attributes. Besides, these also consider discarding the data attributes that provide less value.

Removing Outliers:

Outliers have a significant impact on statistical conclusions, altering and misleading the final analytic result. The dataset was analysed with boxplot, and outliers were detected and deleted. Items with prices greater than 1000000 and sales greater than 1001 were removed.

- item_cnt_day & item_price have extreme values - probably outliers.
- item_price has negative value which doesn't make sense -- probably incorrect data.
- item_cnt_day is negative - suggest return of item. - will remove these entries.

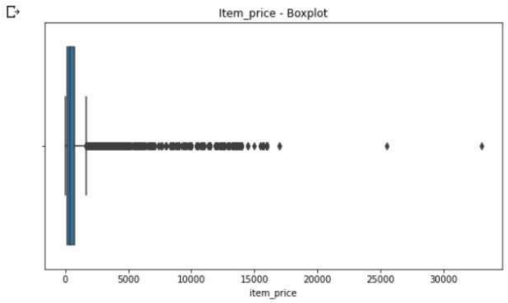


Figure 3 – Boxplot for Item price

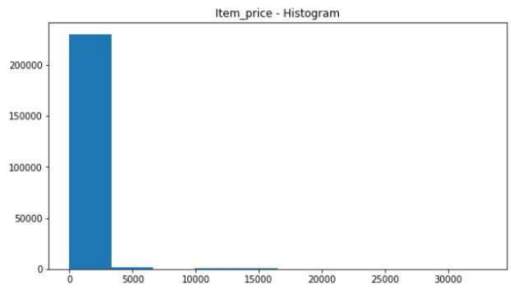
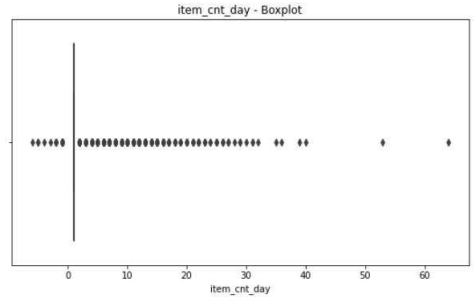


Figure 4 – Histogram for Item Price

The figure 3 shows the boxplot for item_price, where it helps us to understand the trend how the sale of the item. In the similar way the figure 4 shows the Histogram of the item_price whereas it shows that when the price of the item varies.

Extreme Values `Item_cnt_day`:

```
[15] fig, ax = plt.subplots(ncols=2,figsize=(20,5))
     sns.boxplot(x='item_cnt_day', data=eda_df,ax=ax[0])
     ax[1].hist(eda_df['item_cnt_day'])
     ax[0].set_title('item_cnt_day - Boxplot')
     ax[1].set_title('item_cnt_day - Histogram')
     plt.show()
```



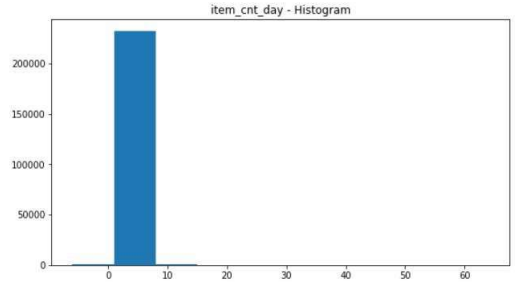


Figure 5- Boxplot for item_cnt_day

Figure 6- Histogram for item_cnt_day

The above figure 5 shows the boxplot for the item sold in the day so here the extreme value is taken as item_cnt_day. Likewise, the figure 6 shows the Histogram of the item_cnt_day whereas it shows that when the count of the item varies in a day.

Exploratory Data Analysis

In statistics, exploratory data analysis is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling and thereby contrasts traditional hypothesis testing.

After data pre-processing, in order to clearly understand the nature of our data, an exploratory analysis was conducted.

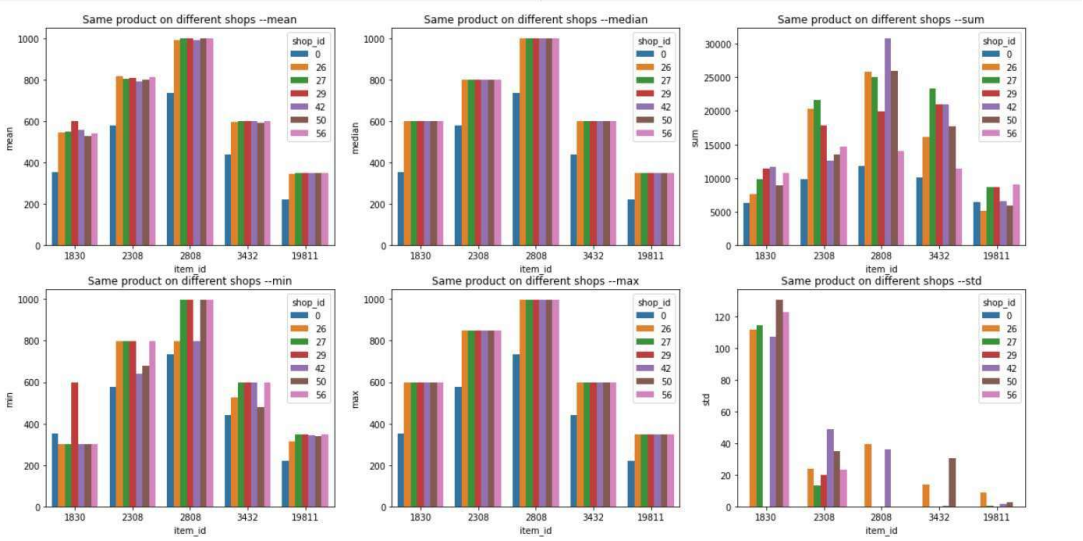


Figure 7 – Item price stat of a product

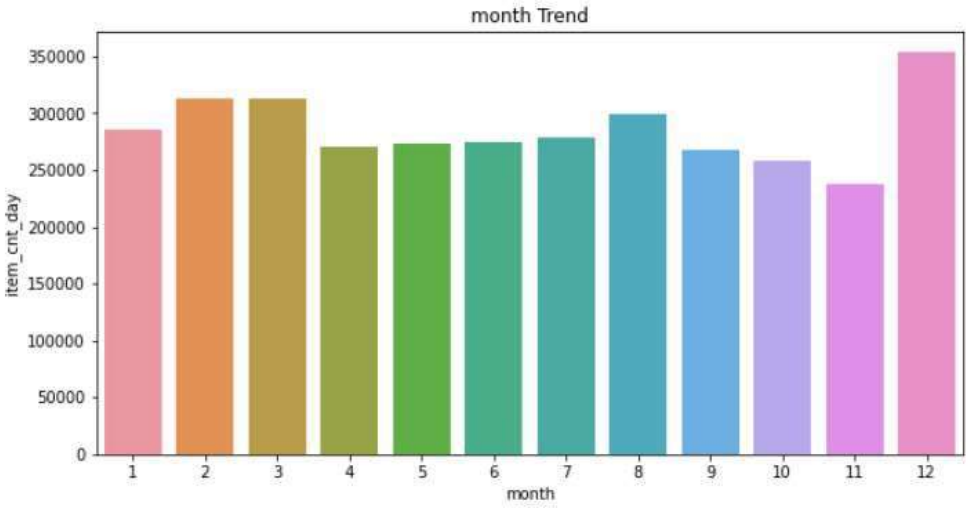


Figure 8 – month vs item_cnt_day

The above figure 8 helps us to understand the item count on the single day by comparing them with the various days in the month which is plotted as histogram for month vs item_cnt_day. It clearly shows that at the month of the 12 the item count was extremely high which is 350000 and similarly the item count is exceptionally low in the month 11 in between 200000 and 250000. And it also shows that it fluctuates in between all the months.

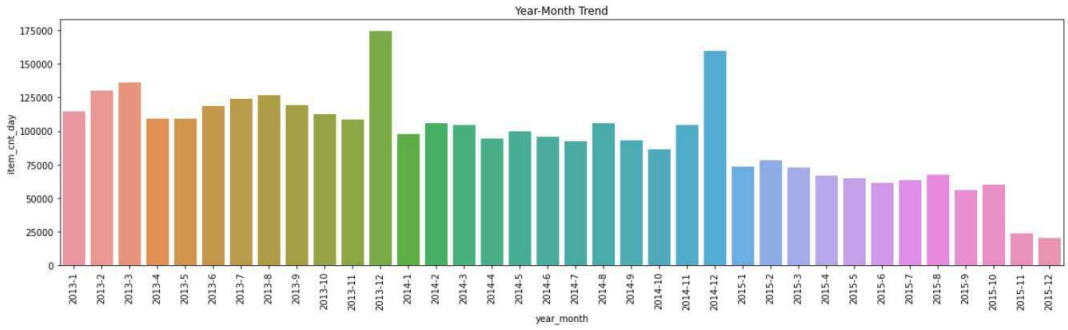


Figure 9 – item_cnt_day vs Year-Month Trend

The above figure 9 shows the graph between the year-month trend vs item_cnt_day. This Histogram depicts the fluctuation of the item count on the yearly and month basis. The item count was exceptionally high in December of 2013 with 175000 items in that specific year and the item count was too low in December of 2015 with minimum of 25000 items in that specific year.

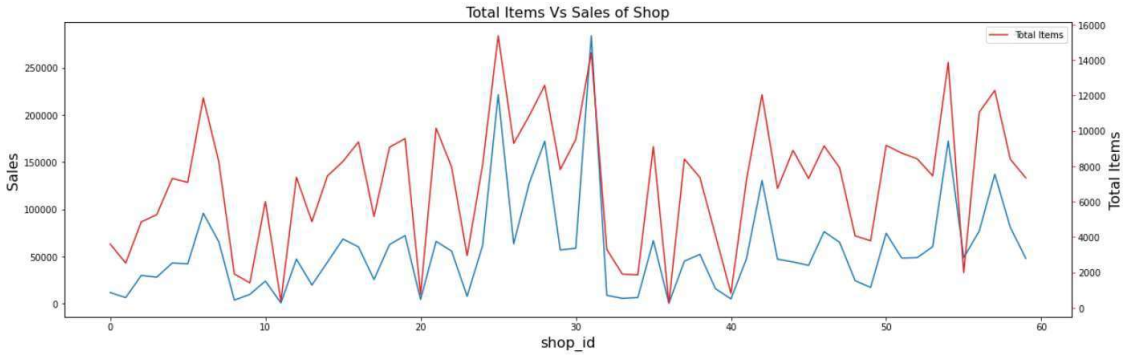


Figure 10- Total Items vs Sales of shop

The above graph shows that the number of items sold in the specific shop which was identified by the shop_id.

Model Building & Evaluation

Here we used 3 types of data to evaluate our model.

Training Data
Testing Data
Validation Data

Training data- This type of data builds up the machine learning algorithm. The model evaluates the data repeatedly to learn more about the data's behavior and then adjusts itself to serve its intended purpose.
Validation data- During training, validation data infuses new data into the model that it hasn't evaluated before.
Test data- After the model is built, testing data once again validates that it can make accurate predictions. Test data provides a final, real-world check of an unseen dataset to confirm that the ML algorithm was trained effectively.

We built a model using the below three algorithms
Linear Regression
Random Forest
XG Boost

Linear Regression
In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values;less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Random Forest

Random forests, also known as random decision forests, is an ensemble learning method for classification, regression, and other problems that works by training a large number of decision trees. For classification tasks, the random forest's output is the class chosen by the majority of trees. The mean or average prediction of the individual trees is returned for regression tasks. Random decision forests address the problem of decision trees overfitting their training set. Random forests outperform decision trees in most cases, however they are less accurate than gradient boosted trees. However, data features can influence how well they function.

XGBoost

Extreme Gradient Boosting (XGBoost) is a distributed gradient-boosted decision tree (GBDT) machine learning toolkit that is scalable. It is the top machine learning package for regression, classification, and ranking tasks, and it includes parallel tree boosting. To understand XGBoost, you must first understand the machine learning ideas and methods on which it is based: supervised machine learning, decision trees, ensemble learning, and gradient boosting.

Supervised machine learning use algorithms to train a model to detect patterns in a dataset with labels and features, and then to predict labels on fresh dataset features using the learned model.

Model Evaluation

The above three methods are used to build the model and now we evaluate using the below technique

There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square
2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)
3. Mean Absolute Error(MAE)

Result:

```
models = [LinearRegression, RandomForestRegressor, XGBRegressor]
models_name = ['LinearRegression', 'RandomForestRegressor', 'XGBRegressor']
model_cache ={}

for i,model in enumerate(models):

    # 01.  train model #
    ml_model = model()
    ml_model.fit(train_X, train_Y)
    model_cache[models_name[i]] = ml_model
    # 02. Predict #
    predictions = ml_model.predict(test_X)
    # 03 Evaluate #
    print('--------------------*--------------------------')
    print('Model: ', str(model).split('_')[-1])
    print('Mean Absolute Error: ',mean_absolute_error(test_Y, predictions))
    print('Explained Variance: ',explained_variance_score(test_Y, predictions))
```

```
--------------------*--------------------------
Model:  base.LinearRegression'>
Mean Absolute Error:  1.2301575139884564
Explained Variance:  0.009925956934689 34
--------------------*--------------------------
Model:  forest.RandomForestRegressor'>
Mean Absolute Error:  0.788392782814444
Explained Variance:  0.73822655515115
--------------------*--------------------------
Model:  <class 'xgboost.sklearn.XGBRegressor'>
Mean Absolute Error:  0.9355150339186218
Explained Variance:  0.6234963473690756
```

Literature Review-

Title: Sales forecasting of Retail Stores Using Machine Learning Techniques.
Authors: A. Krishna, A. V, A. Aich and C. Hegde
link: **https://ieeexplore.ieee.org/abstract/document/8768765**
Summary:

- predict the sales of a retail store using different machine learning techniques
- Trying to determine the best algorithm (GradientBoost algorithm, AdaBoost) suited to our particular problem statement.
- Implemented normal regression techniques and as well as boosting techniques in the approach and have found that the boosting algorithms have better results than the regular regression algorithms.

Title : Machine Learning Model for Sales Forecasting by Using XGBoost
Authors: Xie dairu, Zhang Shilong
Link: https://ieeexplore.ieee.org/document/9342304
Summary:

- We try to discover knowledge of market sales from statistical sales data during the past days and forecast the sales of the the stores.
- Compared to other ensemble learning methods, XGBoost runs ten times faster while uses far fewer resources.
- In the proposed method, since XGBoost is sensitive to outliers, we first do data preprocessing to filter outliers, then convert them into float data type for saving memory. Then aggregated time feature is extracted from different time periods of sales data from day, week and month to year.
- Following that is to capture a variety of features including price feature, rolling feature, lag feature and other statistical features.
- Finally, feature selection is applied to keep those highly relative features for predicating. To effectively evaluate our algorithm, we extensively conduct various experiments on the public Kaggle competition dataset, which contains sales data.

Title: Intelligent Sales Prediction Using Machine Learning Techniques
Authors : Sunitha Cheriyan ,Shaniba ,Saju Mohanan ,Susan Treesa

Link: https://ieeexplore.ieee.org/document/8659115
Summary:

- Finding out the reliable sales trend prediction mechanism which is implemented by using data mining techniques to achieve the best possible revenue.
- Clustering techniques are very useful in discovering distribution patterns and clustering algorithms employ a distance metric-based similarity measures.
- In this project we have performed sales forecasting for stores using different data mining techniques. The task involved predicting the sales on any given day at any store, in order to familiarize ourselves with the task we have studied previously

Conclusion

The purpose of this project was to create a successful model that could forecast future sales for a certain business. Predicting the future will assist the company in precisely estimating costs and revenue, allowing them to forecast short- and long-term success. XGBoost performs better than the other two algorithms. XGBoost technique is decision tree based, imbalanced dataset, outliers and scaling of data is handled very smoothly.

References

[1] Huang, Q., & Zhou, F. (2017, March). Research on retailer data clustering algorithm based on spark. In AIP Conference Proceedings (Vol. 1820, No. 1, p. 080022). AIP Publishing.
[2] Saylı, A., Ozturk, I., & Ustunel, M. (2016). Brand loyalty analysis system using K- Means algorithm. Journal of Engineering Technology and Applied Sciences, 1(3).
[3] Maingi, M. N. A Survey on the Clustering Algorithms in Sales Data Mining.
[4] Sastry, S. H., Babu, P., & Prasada, M. S. (2013). Analysis & Prediction of Sales Data in SAP-ERP System using Clustering Algorithms. arXiv preprint arXiv:1312.2678.
[5] Shrivastava, V., & Arya, N. (2012). A study of various clustering algorithms on retail sales data. Int. J. Comput. Commun. Netw, 1(2).
[6] Rajagopal, D. (2011). Customer data clustering using data mining technique. arXiv preprint arXiv:1112.2663.

[7] Tsai, C. F., Wu, H. C., & Tsai, C. W. (2002). A new data clustering approach for data mining in large databases. In Parallel Architectures, Algorithms and Networks, 2002. I-SPAN'02. Proceedings. International Symposium on (pp. 315-320). IEEE.
[8] Mann, A. K., & Kaur, N. (2013). Review paper on clustering techniques. Global Journal of Computer Science and Technology.
[9] Shah, N., Solanki, M., Tambe, A., & Dhangar, D. Sales Prediction Using Effective Mining Techniques.
[10] Korolev, M., & Ruegg, K. (2015). Gradient Boosted Trees to Predict Store Sales.
[11] Jain, A., Menon, M. N., & Chandra, S. Sales Forecasting for Retail Chains.
[12] Friedman J H. Stochastic gradient boosting[J]. Computational statistics & data analysis, 2002, 38(4): 367-378.
[13] Torlay L , Perrone-Bertolotti M , Thomas E , et al. Machine learningXGBoost analysis of language networks to classify patients with epilepsy[J]. Brain Informatics, 2017.
[14] Ji X , Tong W , Liu Z , et al. Five-Feature Model for Developing the Classifier for Synergistic vs. Antagonistic Drug Combinations Built by XGBoost[J]. Frontiers in Genetics, 2013, 10.
[15] Yin Y , Sun Y , Zhao F , et al. Improved XGBoost model based on genetic algorithm[J]. International Journal of Computer Applications in Technology, 2020, 62(3):240.
[16] Ren X , Guo H , Li S , et al. A Novel Image Classification Method with CNN-XGBoost Model[C]// International Workshop on Digital Watermarking. 2017.
[17] Zhang D, Qian L, Mao B, et al. A data-driven design for fault detection of wind turbines using random forests and XGboost[J]. IEEE Access, 2018, 6: 21020-21031.
[18] Gumus M, Kiran M S. Crude oil price forecasting using XGBoost[C]//2017 International Conference on Computer Science and Engineering (UBMK). IEEE, 2017:1100-1103.
[19] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.
[20] Zhong J, Sun Y, Peng W, et al. XGBFEMF: An XGBoost-based framework foressential protein prediction[J]. IEEE Transactions on NanoBioscience, 2018, 17(3):243- 250.