**Comparative sales analysis of different stores, customers and demographics**

Akash Shukla, Ujjwal Gautam

Department of Computer Science

Lovely Professional University 144111, India

ujjwal97@hotmail.com

ABSTRACT

The era belongs to the internet and the repercussions are such that the everyday data created is enormous. It makes nearly impossible to be processed by man alone. There should be a regular analysis of stores and supermarkets, approver analysis of their sales and customers demand in the sales demography. The total production in the world is totally based on the analysis of the market. The present scenario is that the total demand decides the production of the items. The analysis of the sales and different shops and products tells the profit and loss of the demand supplier. Also, the purchasing power of the customer. This paper elaborates the proper study and analysis of the different stores, customer, demographics, and prediction of the market situation and the visualization will make it easy and will provide the proper insight of the business. The paper is adjoint with project. The project will provide the proper visualization to the data. The sale s and trade data determine that whether we are meeting to the present demand. The data is collected from some outsources.

In this research paper a machine learning model will be trained and will further provide the proper analysis of the data in a detailed manner. This paper will provide the judgmental confidence of setting up new business in field and forecasting the future of the business.

I. INTRODUCTION

During the past twenty years retailers have had to contend with a lot of more subtle and strict customers, new and infrequently unexpected competition from each domestic and foreign sources and a wave of recent technological advances. These and different developments exert continuous pressure on retailers to seek out new and innovative ways in which to differentiate themselves from competitors and adapt to changing and fast environmental circumstances (Dabholkar, Jim Thorpe &amp; Rentz, 1996:3)[1]. The paper is predicated on the sales prediction of the stores. It becomes important for the start-ups as they're unaccustomed market and don't shrewdness seemingly their business model is to carry within the apace flowing business stream. Moreover, the stores that are acting dangerous, will improvise with facilitate of this.

This paper contains the proper and detailed discussion of the analysis of the trade and sales of the shops, as discussed that the visualized part is totally depends on the data provided, the machine learning model involve in the training is Random Forest Classifier. The data contains analysis on all the days, month, weeks, holidays, mostly provide a detailed data on Sunday as the holidays provide the most funding and the supply most of the demand. The name of data is Rossman-store-sales. This paper is also providing a reference to the investors that the seasons on which the investment to be done[2].

The datasets contain 17 fields of study and all the constrains applied on to it. The columns are, stores, Dayofweek, date, year, month, open, promo, state holiday, type, assortment, competition distance, competitionopenscince month, competition opensinceyear, promo2, promoscienceweek, promosinceyear. This paper declares the efficient analysis of the market, sales and the demand of the items in the market and the special strategy of the seller through which the sales cane be incremented. Machine is only the most efficient technique to analyse and make a proper study of the present problem with the given datasets. Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. We are training the model and providing the graphical analysis of the data.

Risk:

1. Mitigates risk: Before planning any business, planning is important but when we plan, we also prepare for the loss in it. Loss of capital and loss of time. This can be managed, or we can take on preventive measure. Market analysis can help, by many ways, it can help in forecasting the future and benefits in the business.
2. Less sale of product: When there is no proper analysis of the market the seller doesn’t aware about the demand and fails to sell the product.
3. Fails to maintain class in market: Whenever the sell is down the image of seller in customers reference descends.

All these risks can only be managed by market analysis, trade and sale analysis. Market study says that the risk is always there, but it can be prevented by market analysis. Market analysis will provide the proper target population by which someone who is going to setup that business area will select different it will save capital as well as time. This is the better way for once to remain safe from potential loss.

Components of store sale and market analysis:

1. Target population
2. Demand of the target population
3. Seasonal sale
4. Holiday special
5. Highest sale day
6. Competitors in the market

These components help to make an analysis[3] that at which time and what demand the supply to be done to raise more revenue.



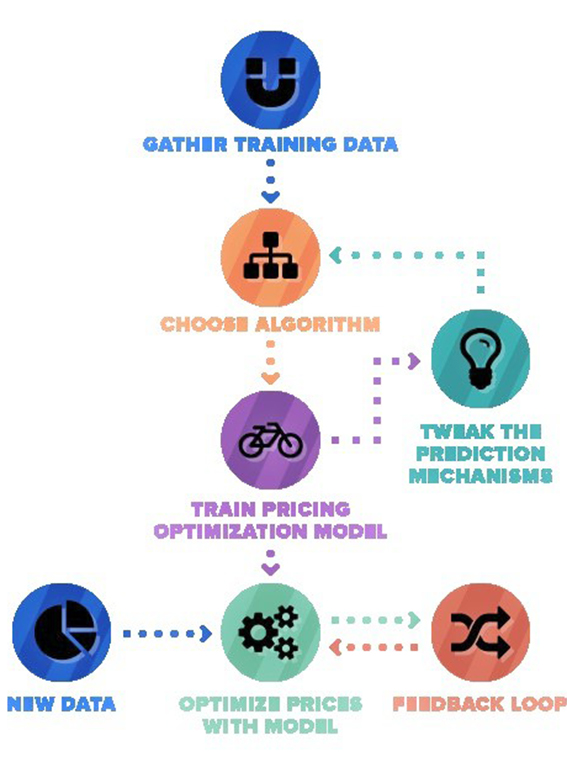
II. MACHINE LEARNING

Artificial intelligence (AI) is that the intelligence of machines and the branch of portable computer technology that ambitions to make it. AI textbooks outline the sphere as "the have a glance at associate degreed layout of intelligent agents" wherever a shrewd agent could be a machine that perceives its surroundings and takes moves that maximize its possibilities of success.

Machine learning is a subfield of artificial intelligence. The aim of the field is to utilize the data into the models. In this recursive procedure, machine continuous updates itself by learning from user interactions[4].

Formally, it is the stated as the ability of machine to learn without being explicitly programmed. It is the tactic of expertise evaluation which automates the analytical model building. It refers to the method by which machines learns the human logic, behaviour patterns and the preferences from the user interaction. The learning from past experiences makes exclusive. It is very tending lately due to its captivating ability to handle complex problems and rigorous improvisation until the goal is met.

The steps involved are as follow:



1. **IMPACT OF MACHINE LEARNING IN RETAILS**

The technology is evolving and, so does the behaviour of consumers. To stay ahead and prevent from the failure of leaning back, the retailers must use the data to study the behaviour of consumers. It is much better than just simple analytics. As compared to which it focuses on the key aspects of customer insights. Moreover, the personalised choices and priorities makes the prediction more precise and accurate. It is the most promising when it comes to niches and sizes. The number of companies adopting it readily is increasing, exponentially. The machine learning algorithms can help in identifying the root cause of various faults within the system.

A global market insight study has shown that the retail sector investment in AI will increase to $8 billion by the year, 2024[5]. It is just because the foreseen rapid pace growth in the sector. Moreover, it will reduce human error and boost the speed within the various segments of the sector. It the necessity of the era for the retailer to adjust within this digital evolution.

Benefits of using machine learning in retail are:

* It helps in delivering niche products.
* It can enhance the income by adjusting the prices in real time.
* The segregation of class of customer can be done better on the prior experiences.
* Better prediction of sales and consumer service based totally upon formerly customer behaviour data.
* It can help in making inventory plans and providing better maintenance with help of accurate and improvised predictions.
* The finding faults in the system had never been easier and more efficient.

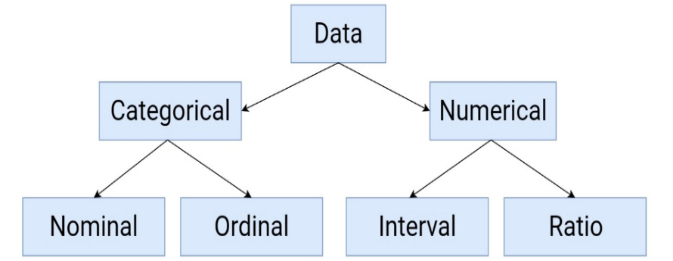
1. **DATA PREPROCESSING**

**To handle the data the first step is data preprocessing. The most important step in data mining, collection of data contains irrelevant sets which is of no use. Data preprocessing is the first filtering step of the data sets. The term ‘garbage in and garbage out’ is mostly used in data preprocessing which means removing irrelevant values and filtering out a precise data sets from the garbage from of data. The data collection steps do not include filtering or arranging data. In machine learning processing of data is the only technique to do correction in data or make data more consistent.**

**The data provided can be viewed as features and the**

**Attributes which should to perfectly handled by data preprocessing.**

**Data contains many from of the sets and values in it.[6]**



**The data processing starts with collecting of data from human resources which contains many errors, in case of large data it is impossible for human to remove error and to processes it. So, we use machine learning models like SVM, Decision Tree, Robert frost etc.**

**The data preprocessing steps includes:**

1. **Data Quality Assessment**
2. **Feature Aggregation**
3. **Feature Sampling**
4. **Dimensionality Reduction**
5. **Feature Encoding**
6. **Training, Validating and testing**

**1.Data Quality Assessment**

**The data collected from many sources which is not that efficient and provide some wrong and incomplete values in to the data provided.**

1. **Missing Values.**
2. **Inconsistence Values.**
3. **Duplicate Values.**

**2. Feature Aggregation**

**Feature aggregation is done to find the most aggregate and most fitted aggregate value to sort the data, and filter the data. The advantage of feature aggregation is to reduce memory consumption to store the data and reduce the processing type. It provides us the high-level visualization of the data over particular object.**

**3. Feature Sampling**

**Sampling means selecting the sample of data on which we are working or training our model. If we don’t divide our data into samples it may be expensive for the system.**

**The best sampling is done using classes, random sampling may give overfitting to the data during training and testing.**

**4. Dimensionality Reduction**

**Reducing dimensionality means reducing the dimensions of data without removing samples which have more dimensions rather removing the constraints which increases the dimensions. If the number of dimensions are more than the visualization is very difficult. As the data sampling increases the dimension increases and that creates the problem in visualization.**

**5. Feature encoding**

**Basically, feature encoding means transformation of the data so that it fits the machine learning model properly and provide the correct classification in the data.**

**There are few types of encoding:**

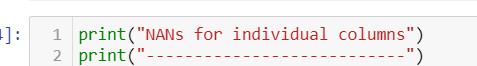
1. **Nominal**
2. **Ordinal**
3. **Interval**
4. **Ratio**

**6. Training, validating and testing**

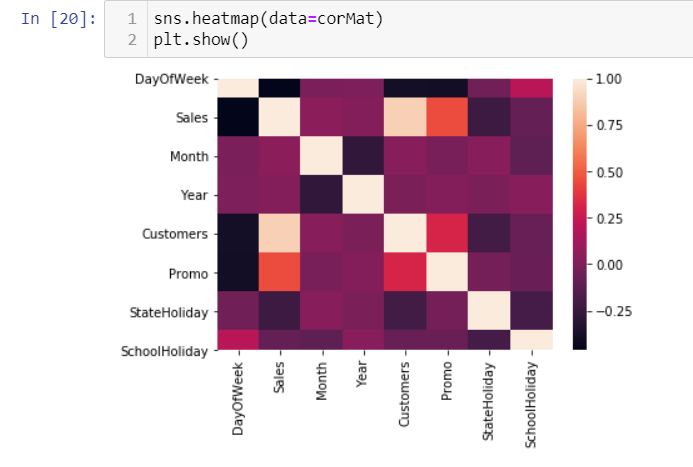
**The most important aspect of the machine learning algorithms, the model is trained on the testing data samples, after the model works well then, the model is tested on some data. To validate that the machine learning system is working and providing the correct analysis[7].**

**Steps involved in the data pre-processing in the corresponding project are: -**

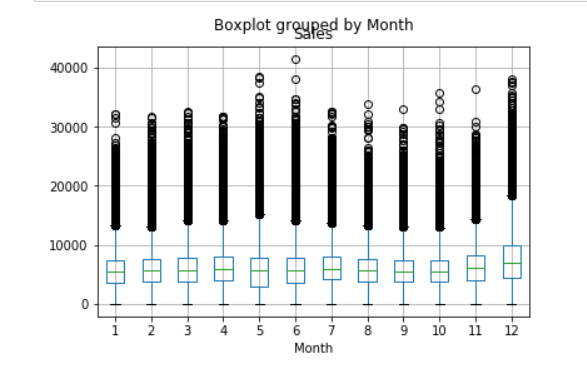
* **Importing datasets with corresponding delimiters and data types.**
* **Splitting the sub-attributes of composite attributes.**
* **Finding the NaN and null values, and using mean to remove it.**

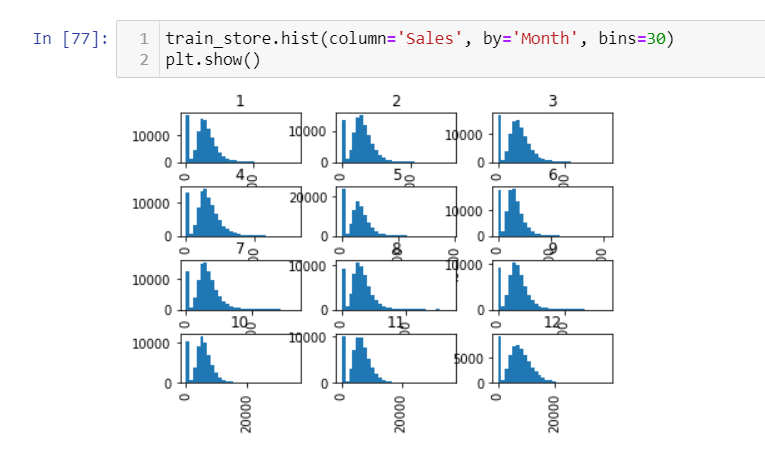


* **Finding the correlation using pandas.corr().**
* **Visualizing it using heatmap**



* **Comparison among train and test datasets, so that the columns overlap.**
* **Verify if any missing value is left**
* **Repeating the same for the store dataset**
* **Finally, merging the train and store dataset into the resultant dataset, which is further sent for the modelling**
* **Visual exploration using boxplot and histogram.**





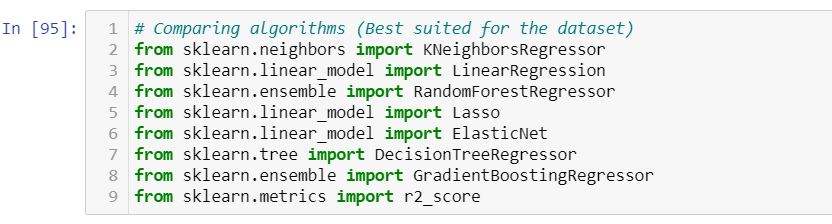
* **Finally, the data sets are ready for modelling.**

1. **MODELLING**

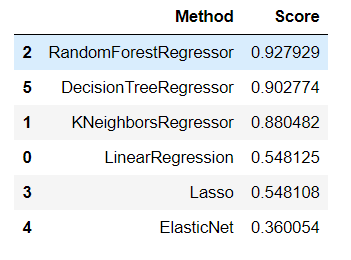
**Ensemble modelling is the most frequently model for uplifting the productivity in cases of sales and churn prevention.**

**The features are categorical, on the other hand the target i.e. sales, is continuous. To handle this, we need such a model which can work for both classification as well as regression. Thus, random forest comes into play. We can completely apply random forest on the stores[8].**

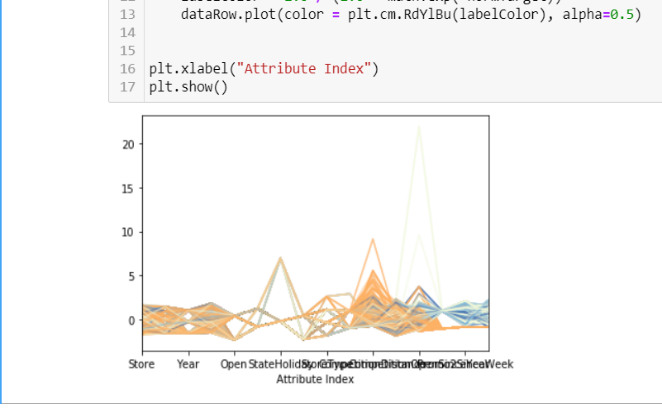
**Firstly, the tuples with sales as 0, are to be removed. Date must be splitted into its sub-attributes. The data of train and store ought to be merged to get the complete display of picture.**



**Bootstrapping is used by random forest for the training. On the other hand, gradient boosting builds decision tree. The resultant prediction is the weighted average.**



**After checking the R2-Score of the various machine learning model[9], the Random Forest landed on the top. Therefore, we have used it. Finally, storing the result into the final\_report inside the result directory.**



1. CONCLUSION

Comparing and making an analysis is the better way of planning and proceeding to the plan. Machine learning provided the better way to make analysis of the departmental stores and trade done by them. The machine learning model used in this project is random forest as the model is most accurate the analysis is shown through many visualizations using matplotlib library in python. We used heat plot to show the effective output of the analysis, the snaps of the data set also shown above just to get clear about the analysis processes of the data in machine learning model. This project and the related report will provide a technique to the market research as in the year 2020 the world is going to suffer a great market crisis. The test error feedbacked from kaggle.com is 0.12742.

1. REFERENCES

[1] C. F. D. Sciences- and F. Project, “Computing For Data Sciences-.”

[2] J. L. B, M. Castrill, and E. Ram, “Video Analytics for Audience Measurement,” vol. 8811, pp. 53–65, 2014, doi: 10.1007/978-3-319-12811-5.

[3] S. Kolassa, “Random Forest Density Forecasts in Retail,” no. June, 2017.

[4] N. V. Razmochaeva and D. M. Klionskiy, “Data presentation and application of machine learning methods for automating retail sales management processes,” *Proc. 2019 IEEE Conf. Russ. Young Res. Electr. Electron. Eng. ElConRus 2019*, no. 1, pp. 1444–1448, 2019, doi: 10.1109/EIConRus.2019.8657077.

[5] S. Elias and S. Singh, “FORECASTING of WALMART SALES using MACHINE LEARNING ALGORITHMS,” 2018.

[6] A. Krishna, V. Akhilesh, A. Aich, and C. Hegde, “Sales-forecasting of Retail Stores using Machine Learning Techniques,” *Proc. 2018 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2018*, pp. 160–166, 2018, doi: 10.1109/CSITSS.2018.8768765.

[7] L. Kirichenko, T. Radivilova, and I. Zinkevich, “Comparative analysis of conversion series forecasting in e-commerce tasks,” *Adv. Intell. Syst. Comput.*, vol. 689, no. 1, pp. 230–242, 2018, doi: 10.1007/978-3-319-70581-1\_16.

[8] D. Beam and M. Schramm, “Rossmann Store Sales,” no. December, pp. 1–6, 2015.

[9] S. Cheriyan, S. Ibrahim, S. Mohanan, and S. Treesa, “Intelligent Sales Prediction Using Machine Learning Techniques,” *Proc. - 2018 Int. Conf. Comput. Electron. Commun. Eng. iCCECE 2018*, pp. 53–58, 2019, doi: 10.1109/iCCECOME.2018.8659115.