Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 7- Issue 1, January 2024
Paper : 112

Sales Forecast Prediction with a Web Interface Using Flask and Machine Learning

Dr. Dhanakoti V #1, Sakthivel S#2, Shakir Ahamed M#3, Sasitharan M#4

[1] *Professor, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Chennai, Tamil Nadu, India*
[2,3,4] *Department of Computer Science and Engineering, SRM Valliammai Engineering College , Chennai, Tamil Nadu, India*

Abstract— **Business intelligence is one of the demanded skills in information technology, as it is the current stipulation. We have developed model where it predicts the future sales pattern of a product by imparting the purchase history between a time period to the model. This project focuses on analyzing and visualizing the regional sales of products. The underlying algorithm is based on the linear regression and the random forest classifier. Segmenting consumer-based buying behavior and applying 80/20 rule to identify top customers/products by applying Xgboost.**

***Keywords*— Business intelligence, Inventory Management, Machine Learning, Time Series Prediction.**

I. Introduction

Most of our buying decisions are not based on well-defined logic. Emotions, trust, communication skills, culture, and intuition play a big role in our buying decisions. Although humans do not follow a well-defined logic, we do have some repeated patterns. We often buy the same things and behave in a similar way. Nowadays, shopping malls and Big Marts keep track of individual item sales data in order to forecast future client demand and adjust inventory management based on the prediction that is done on that data using various machine learning algorithms.

II. Related Works

There are various machine learning algorithms that can be used in the sales forecasting. Orogun Adebola et al. explains that Predicting customer behavior is an uncertain and difficult task. Thus, developing customer behavior models requires the right technique and approach. Despite the complexity of this formulation, most customer models are relatively simple. Because of this necessity, most customer behavior models ignore so many pertinent factors that the predictions they generate are generally not very reliable [1]. Gyanendra Chaubey et al. proposes that this paper presents a comparative study of different machine learning techniques that have been applied to the problem of customer purchasing behavior prediction. Experiments are done using logistic regression, decision tree, k-nearest neighbors (KNN), Naïve Bayes, SVM, random forest, stochastic gradient descent (SGD), ANN, AdaBoost, and dummy classifier, as well as some hybrid algorithms that use stacking like SvmAda, RfAda, and KnnSgd. Furthermore, the confusion matrix and ROC curve are used to calculate the accuracy of each model [2]. B. Arivazhagan et al. proposes that this paper examines E-Commerce datasets in order to discover useful and interesting patterns by employing data mining association rules to achieve the best results. The rules were generated by the FP-Growth algorithm from frequently used item sets. The research is implemented in the rapid miner tool and evaluated using appropriate evaluation metrics [3]. K V.L Sita Ratnam et al. discusses the competitive world is searching for customer-oriented approach. The requirements of a customer are collected in the form of data which is to be analyzed for obtaining customer behavior. Therefore, inquiry is made to determine the characteristic features of a customer by using the data which is collected from different sources of all the online shopping sites [4]. Adil Mahmud Choudhury et al. states that to lift the revenue boundary and stay ahead of the competitors it is important to understand customer's purchase behavior. Different business industries proposed different policies to explore the potentiality of a customer based on statistical analysis. In this paper, we rather propose a machine learning approach to identify potential customers for a retail superstore. The paper proposed an engineered approach to classify potential customer, based on previously recorded purchase behavior [5]. Quang Hung Do et al. proposes that the main goal of this study is to investigate the classification capability of several machine learning (ML) techniques, including decision tree (DT), multilayer perceptron (MLP) network, Naïve Bayes, radial basis function (RBF) network, and support vector machine (SVM) for predicting purchase decisions. The application case is related to consumer purchase decisions of domestic goods in the context of

Vietnam. Firstly, factors influencing Vietnamese consumers' purchase decision of domestic products were identified. Then, data from 240 consumers in Vietnam were collected [6]. Pornpimon Kachamas et al. states that Artificial intelligent model was developed by the results from 75 specialists who evaluated the behavior that will likely occur after the comments have been posted. The results, hence, were collected and prepared for the data modelling process using the Naïve Bayes probability concept, afterwards, testing for the model's accuracy with 10-fold cross validation technique [7]. Bo Zhao et al. states that it is important for merchants to identify who can be converted into repeated buyers. By targeting on these potential loyal customers, merchants can greatly reduce the promotion cost and enhance the return on investment (ROI). It is well known that in the field of online advertising, customer targeting is extremely challenging, especially for fresh buyers. With the long-term user behavior log accumulated by Tmall.com, we get a set of merchants and their corresponding new buyers acquired during the promotion on the "Double 11" day [8]. Saifil Momin et al. states that this chapter comprises an experimental comparison of various traditional classification algorithms, namely K-nearest neighbors, naive Bayes, random forest, decision tree, and logistic regression, with artificial neural network to predict the customer churn on IBM's Telco Customer Churn dataset. We have compared these models based on their accuracy in predicting customer churn. Our ANN model achieves an accuracy score of 82.83% on validation data, better than our performance of 79.86% achieved for the traditional approach of using K-nearest neighbors [9]. Rana Alaa El-Deen Ahmeda et al. states that in this paper eleven data mining classification techniques will be comparatively tested to find the best classifier fit for consumer online shopping attitudes and behavior according to obtained dataset for big agency of online shopping, the results shows that decision table classifier and filtered classifier gives the highest accuracy and the lowest accuracy is achieved by classification via clustering and simple cart, also this paper will provide a recommender system based on decision table classifier helping the customer to find the products he/she is searching for in some ecommerce web sites [10].

III. Materials And Methods

A. Various library packages are available to process text, to split the data into training and testing, and to feed the training data to the algorithms. Finally, some visualisation plots and performance metrics such as accuracy and losses are presented.

1) Train Test Split: One of the most important parts of machine learning programs is data. How we compute and evaluate it matters a lot. Sci-kit Learn has a method where it splits the data into training and testing data. Training data is something we feed into the proposed model to learn a variety of parameters, including patterns, outliers, and insights. Testing data is provided by the users to check the model's performance, like accuracy and loss. It totally depends on the performance of the model.

2) Accuracy Score: The most significant factor in the model is accuracy. To achieve good accuracy, the model must have been properly manipulated. Datasets, model architecture, and finally, the methods and modules that are used in it. The more accuracy the model provides, the more it is considered top-notch performance.

3) Label Encoding: In machine learning, we usually deal with datasets that contain multiple labels in one or more columns. These labels can be in the form of words or numbers. To make the data understandable or in a human-readable form, the training data is often labelled in words. Label encoding refers to converting the labels into a numeric form to make them machine-readable. Machine learning algorithms can then better decide how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

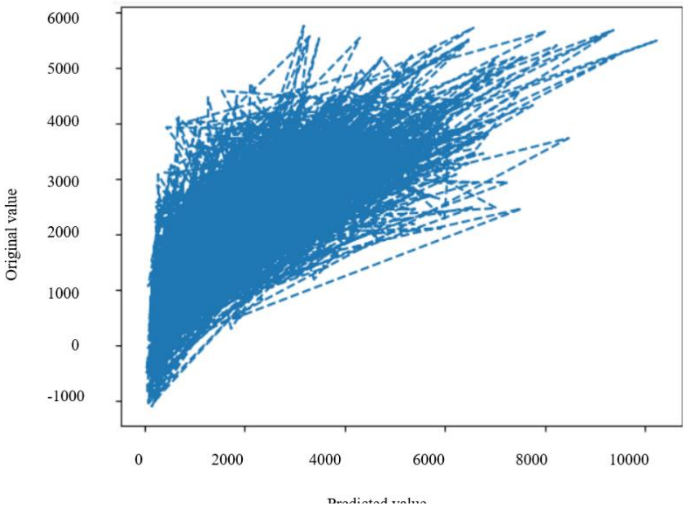
4) Standardization: Standardization is one of the feature scaling techniques that scales down the data in such a way that the algorithms (like KNN, Logistic Regression, etc.) that are dependent on distance and weights should not be affected by uneven-scaled datasets because if it happens, then the model accuracy will not be good (will show this practically). On the other hand, if we scale the data evenly in such a way that the data points are mean-centric and the standard deviation of the distribution is 1, then the weights will be treated equally by the algorithm, giving more relevant and accurate results.

5) Joblib: The machine learning library scikit-learn also uses joblib behind the scenes for running its algorithms in parallel (scikit-learn parallel run info link). Joblib is basically a wrapper library that uses other libraries for running code in parallel. It also lets us choose between multi-threading and multi-processing.

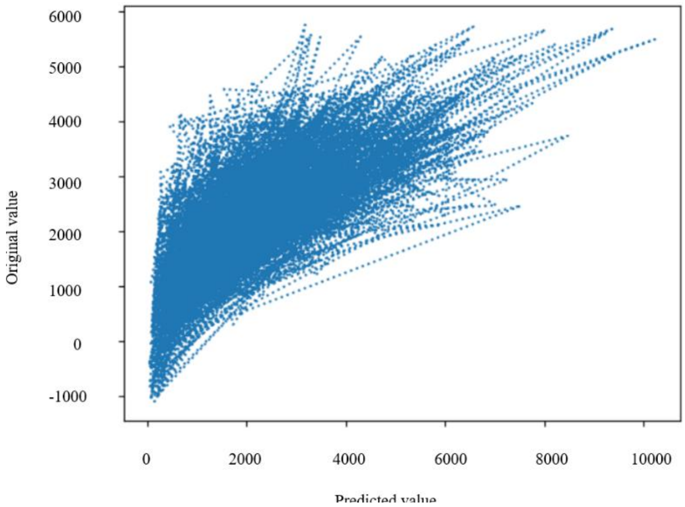
6) R2 Score: As this is a multiclass/label prediction model, the regression approach is used. For a classification problem, it is a conventional and apt method to compute accuracy score, but as this is a regression model, we compute r2Score (coefficient of determination), in short terms, it is the preposition of difference between the expected value and the predicted value. Regression does not have the accurate value to predict; rather, the performance is computed as a value that is closer to the true value.

B. Algorithms

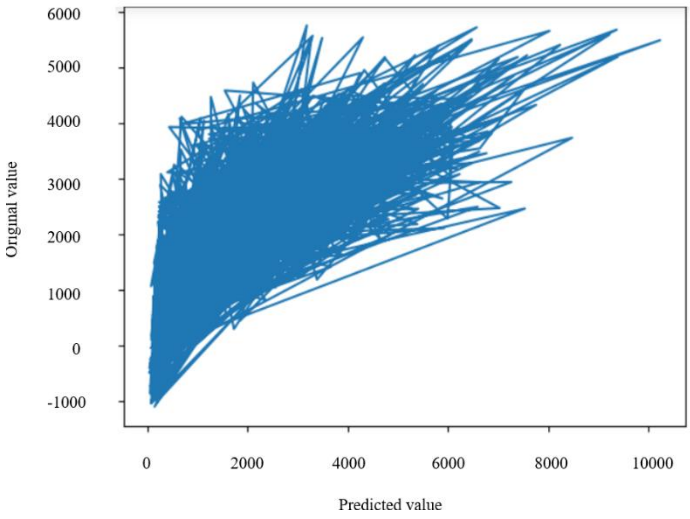
1) Linear Regression: Linear regression is a machine learning algorithm based on supervised learning. This is one of the most widely used algorithms that can be used to solve regression problems. It performs a regression task. Regression models predict a target value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables they are considering and the number of independent variables they use. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressor. The independent variables can be called exogenous variables, predictor variables, or regressors. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change. The mathematical functionalities behind them are based on two major factors: the slope and the linear boundaries that separate the labels. As this is a regression model, the r2 score (coefficient of determination) has been calculated. The r2 score of the linear regression algorithm is 0.500721.



2) Random forest classifier: Random Forest is a popular machine-learning algorithm that belongs to the supervised learning technique. It can be used for both classification and regression problems in ML. It is based on the concept of ensemble learning, which is the process of combining multiple classifiers to solve a complex problem and improve the performance of the model. The r2 score of the Random Forest algorithm is 0.52734.



3) XGBoost Algorithm: XGBoost is a robust machine-learning algorithm that can help you understand your data and make better decisions. XGBoost is an implementation of gradient-boosting decision trees. It has been used by data scientists and researchers worldwide to optimize their machine-learning models. XGBoost stands for "Extreme Gradient Boosting". XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. It uses parallel tree boosting to solve many data science problems in a fast and accurate way. The r2 score of the XGBoost algorithm is 0.53361



IV. Implementation

For any project to begin, there must be a user-friendly IDE (integrated development environment) to begin with. Here is a usage for 3IDE's in this project: Atom is used for the front-end development. The reason for this is that it is open source, provides better syntax highlighting, has strong community support, and has various plug-ins that users can download and interact with. Jupyter Notebook has been used for the back-end application, which is the machine learning model. Although it has a certain number of demerits, the main advantage is that it is kernal-based, in which the user has the flexibility to work on a particular kernal in an "isolated manner". PyCharm has been used for server-side computing, which is the middleware that makes an established connection between the Atom and Jupiter Notebook. The server is made to run on a local host at first, but after all the consulting, it has been hosted on the internet. The project has been fractionated into 4 modules, the first and foremost of which is the front-end (Web Interface), which the users see and interact with. Several languages and frameworks have been used, including HTML, CSS, and JavaScript. An important aspect of the project is the use of Content Delivery Network (CDN), which speeds up the connection of Bootstrap. There have been several GIFs used to make the interface look aesthetic. The inclusion of the bootstrap framework simplifies the micro-static works such as padding and margins around HTML tags and makes the interface mobile compatible, which is the main merit of bootstrap. Now coming to the middleware, there is a usage of the Flask framework where it is used to extract values from the interface and forward it for further computation. There is a usage of a function named "result" to extract values and store them in a variable, and then with that variable, prediction happens depending on the user inputs and two more function has been used one is to create table namely "create table" and to insert user inputted values namely "insert row," The above-mentioned functions are totally based on creating databases and inserting the user-inputted values. The predicted value is displayed in the web interface by using a method called "render template". Before going into the database section, let us jump into the back-end (ML Model). Various Python packages are used, some of which are: pandas for reading the training data frame; dtale and pandas profiling for data visualization on various formats; label encoder for converting categorical values to numerical values; train test split for splitting the data by the 80/20 rule; standardscaler for standardizing the values; joblib for saving the model; and three algorithms (linear regression, random forest regression, and XGBoost) for prediction. And finally, the database, which contains the attribute values of the user inputs, can be viewed using the DB SQLite viewer. as mentioned above, the two functions help perform and complete the task. After various research, the project is hosted through a feasible hosting service provider so that users can notice it on the World Wide Web.

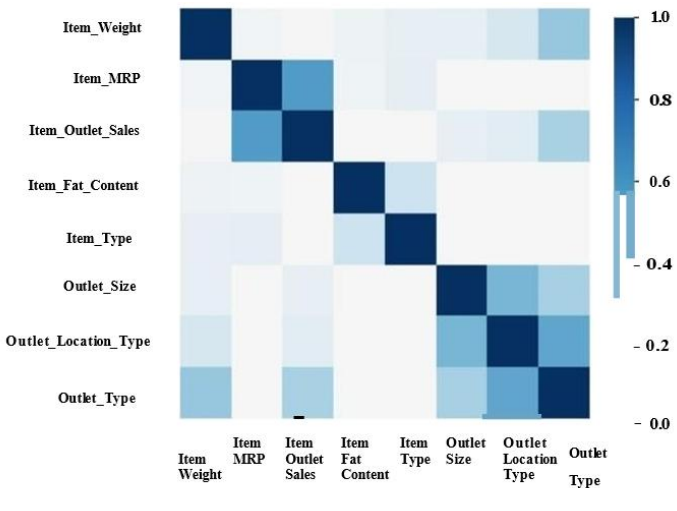


Fig. 4. 1

Fig.4.1 Here is a correlation graph of eight attributes,a heatmap, where it represents the dependency of each attribute or feature column present in the dataframe. The correlation values lie between 0 and 1, where 0 is the least related and 1 is the most related. Here, the darkest color represents the highest correlated value between two columns, and the lightest color represents the lowest correlated value between two columns. Let us delve deeper into the graph,there is a color range scale on the right side of the graph that tells the degree of correlation,every column to its same has the highest correlation value,thus it is dark in color and has a value of 1,for example, there is a

diagonal dark line in the graph that depicts the same attributes,"Item Outlet sales" and "Item MRP" the graph depicts 0.8,the more Maximum retail price of an item will generate more Outlet Sales,The least correlated value is 0, outlet type and item fat content are definitely not dependent on each other, and item type and outlet type also fall under this category.

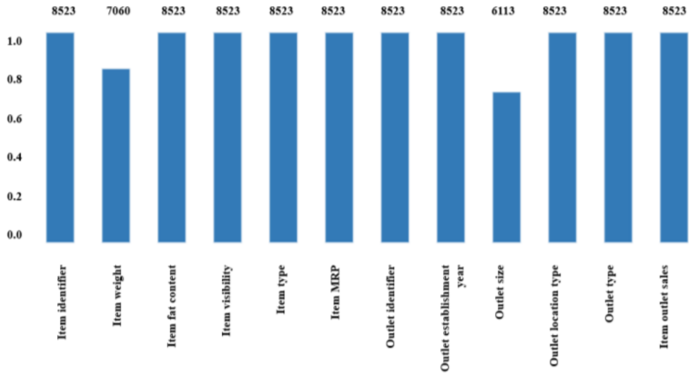


Fig. 4. 2

Fig.4.2 represents that before data preprocessing,every dataframe has some missing or disappeared values.A pictorial representation of missing values in each feature columns using bar graph. For a developer to follow and go through the process of SDLC (software development life cycle), the fundamental step is to collect the data and information and gather domain knowledge. No source would provide a 100% suitable dataset to proceed further; thus,the developer has the responsibility to pre-process the dataset, such as standardizing,filling in the missing values,converting the categorical values to numerical values (label encoding), and more. Standardizing is nothing but converting the data format into a single, common format so that ML algorithms can easily manipulate it. There are different procedures to fill up the missing values; some of them are mean, median,Mode.The main advantage is that we get more data. The more data a prediction system has,the higher its accuracy and performance it outcomes.The dataset also has categorical values; it won't be suitable for prediction, so there is a need for conversion of categorical values to numerical values so that ML algorithms can be applied for prediction. Here is the pictorial representation of missing values in each column. A total of 8523 observations have been made, and as the representation seeks to exhibit the missing values in probability, the total observations have been divided into five,thus the initial value in the graph starts with 1704 and added up consecutively.Item weight has 7060 values out of 8523, and outlet size has 6113 values out of 8523.

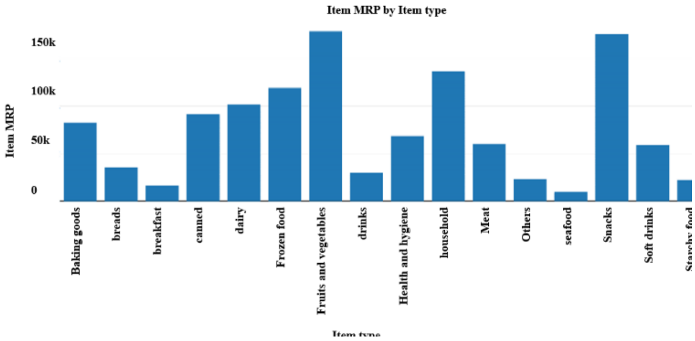


Fig. 4. 3

Fig 4.3 For a data scientist, it is important to know the insights and patterns and must gather as much information as he can. Here is the attribute comparison graph. It is important to compare the attributes and observe the fluctuations, and this graph reveals the MRP of a particular product. on the Y-axis to Product type (Diary foods, Breakfast, Breads, etc.) on the X-axis. The graph shows that fruits, vegetables, and snacks are the most expensive, while seafood is the least expensive. The Y-axis and X-axis are separated by the total price of each product, which ranges from 0K to 150K.The second most expensive product is household commodities, which cumulate up to 140K, the third is frozen foods, which sum up to 125K, and the remaining products and their computations This graph also projects peripheral knowledge such as demanded products, availability of a product, and the cost of the raw materials required to manufacture the final product (highly priced raw materials lead to an expensive finished product).

V. Discussion

Starting from the environment being coded to the final output system projects, everything is done in Python majorly and web-oriented languages. The Jupyter notebook is used as an integrated development environment as it facilitates kernels that will be useful for machine learning and deep learning. And coming to the part of coding, there is the usage of various library packages, which include Matplotlib, Pandas, and Scikit Learn. Furthermore, the data is split into training and testing derived from a method model selection with a testing size of 20% of the data we feed, and after that, various algorithms are applied to check their performance by visualizing it, comparing its accuracy, and finally predicting outlet sales.

VI. Conclusion

Although there are various methods that have been suggested for detecting future sales. Our study is the first of its kind to employ machine learning to analyse consumer purchasing patterns for a major retail chain. To record the connection between categories, items, amount, measurement unit, and sales, we created features Correctly recognizing a potential consumer can greatly help a firm. A tailored marketing strategy can be used to address the potential customer, increasing a business's sales. In the future, machine learning can be used to identify client behaviour, product interest, and purchasing frequency, allowing for more appropriate marketing plans and effective supply chain management.

References

[1] Orogun Adebola and Bukola Onyekwelu, Predicting Consumer Behaviour in Digital Market: A Machine Learning Approach, 2019, Pp 17-23.

[2] Gyanendra Chaubey, Prathamesh Rajendra Gavhane, Dhananjay Bisen, Siddhartha Kumar Arjaria et al., Customer purchasing behavior prediction using machine learning classification techniques, Journal of Ambient Intelligence and Humanized Computing, 2022, Pp.25-31.

[3] B. Arivazhagan, S. Pandikumar, S. Bharani Sethupandian, R. Shankara Subramanian et al., Pattern Discovery and Analysis of Customer Buying

Behavior Using Association Rules Mining Algorithm in E-Commerce, First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, Pp.19-24.

[4] Dr M.R. Narasingha Rao, K V. LSita Ratnam, M D.S. Prasanth, P. Lakshmi Bhavani et al., A Survey on Analysis of Online Consumer Behaviour Using Association Rules, International Journal of Engineering & Technology, 2018, Pp.36-40.

[5] Adil Mahmud Choudhury, Kamruddin Nur, A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior, International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2019, Pp.26-31.

[6] Quang Hung Do, Tran Van Trang, an approach based on machine learning techniques for forecasting Vietnamese consumers' purchase behaviour, 2020, Pp.14-19

[7] Pornpimon Kachamas, Suphamongkol Akkaradamrongrat, Sukree Sinthupinyo, Achara Chandrachai et al., Application of Artificial Intelligent in the Prediction of Consumer Behavior from Facebook Posts Analysis, 2019, Pp.37-41.

[8] Bo Zhao, Atsuhiro Takasu, Ramin Yahyapour, Xiaoming Fu, Loyal Consumers or One-Time Deal Hunters: Repeat Buyer Prediction for E-Commerce, 2019 International Conference on Data Mining Workshops (ICDMW), 2019, Pp.24-29.

[9] Saifil Momin, Tanuj Bohra, Purva Raut et al., Prediction of Customer Churn Using Machine Learning, EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing, 2020, Pp.19-26.

[10] Rana Alaa El-Deen Ahmeda, Mohamed Elemam, Shereen Morsya, Nermeen Mekawiea et al., Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behavior Using Data Mining, Fifth International Conference on Communication Systems and Network Technologies (CSNT), 2019, Pp.27-36.