Grenze International Journal of Engineering and Technology, June Issue

Automated Product Categorization and Sales Forecasting: A Machine Learning Approach for Retail Analytics

G. Harshini[1], T. Thrisha Reddy[2], T Prathima[3] and A Sirisha[4]
[1-2]Department of Artificial Intelligence and Data Science, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana,
Email: {harshinigande, thrishathondapu}@gmail.com
[3-4]Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana,
Email: {prathimareddy.t, asirishanaidu}@gmail.com

***Abstract*— Due to new technologies, sales forecasting has become increasingly popular as a way to improve market operations and productivity in the retail industry. Although the business has historically concentrated on a stand-ard statistical model, machine-learning techniques have garnered increased attention in recent years. A key com-ponent of retail is store sales. The most trustworthy models are scrutinized by administrators in order to help fore-cast future sales. Making decisions based on historical and present data might be facilitated by forecasting future variations or increases in store sales. By recognizing sales patterns and trends, accurate forecasting will help busi-nesses or retailers increase profitability and enhance the customer experience. Predicting retail store sales with deep learning and machine learning (ML) approaches yields high. This study summarizes ML techniques to identify the best algorithm for predicting retail sales based on a collection of prior research articles. This work detects the most influencing attributes that affect sales, and suitable ML algorithms for sales forecasting.**

***Index Terms*— Ridge regression, Lasso, Random Forest regressor, product categorization, Sales prediction**

I. INTRODUCTION

In the past, businesses would manufacture things regardless of demand or sales volume. Data about the mar-ket's demand for products is necessary for any producer to decide whether to raise or decrease the production of multiple units. Businesses that disregard these principles when competing in the marketplace risk losing out. Various businesses select particular standards to ascertain their sales and demand[1]. Accurate and time-ly forecasting of future revenue, also known as revenue forecasting or sales forecasting, can provide im-portant insight to organizations involved in the manufacture, distribution, or retail of goods in today's high-ly competitive economy and ever-changing consumer landscape [2]. Long-term predictions can deal with corporate growth, decision-making and business growth, while short-term forecasts are primarily helpful for production planning and stock management [1].
Because many products in industries have a limited shelf life and can result in income loss in both surplus and

Grenze ID: 01.GIJET.10.2.539_3

shortage circumstances, sales forecasting is especially crucial. A shortage of products results from plac-ing too many orders, and a lack of opportunities arises from placing too few orders. Because of this, market rivalry is constantly changing as a result of elements including price, advertising, and rising consumer de-mand [3]. Sales projections are typically made at random by managers. On the other hand, professional man-agers become scarce and unpredictable. Computer programs that can substitute for qualified managers when they are unavailable or give prospective sales forecasts can help with sales predictions and help decision-makers make the best choice.
Attempting to simulate the abilities of professional managers within a computer program is one way to put such a strategy into practice [4]. As an alternative, ML techniques can be utilized to automatically create reliable sales forecasting models by utilizing the wealth of sales data and related information. This method is far more straightforward. It is flexible enough to adjust to changes in data and is not influenced by the unique characteristics of a single sales manager. It may, however, overstate the precision of a human expert's prognosis, which is often imprecise. For instance, in the past, businesses produced goods regardless of demand or sales volume since they were experiencing a lot of issues. Any producer must have data on consumer demand for items in order to determine whether to increase or decrease the number of units being produced, as they are unable to determine how much to sell. Businesses will suffer losses if they compete in the market without taking these concepts into account. Various businesses select various criteria to ascertain their market and sales.
Businesses have historically relied on a variety of statistical models, including time series and linear regression, feature engineering, and random forest models, to anticipate future sales and demand. These are just a few methods for sales forecasting. A time series is made up of data points that are kept for a certain amount of time and are used to predict future events. A time series is a set of data points that are gathered over a period of time at consecutive, equally spaced intervals. The most crucial elements to examine are cyclicity, irregularity, seasonality, and patterns. A mathematical technique called linear regression is used to project historical values. It can assist in identifying the underlying patterns and resolving situations where rates are exaggerated[5, 6]. Feature engineering is the process of creating features using domain knowledge data in order to improve the accuracy of predictive ML models. It results in more insightful data analysis and practical viewpoint[7].

A. Objectives

- Using ML techniques, the products will be automatically classified based on their product descriptions.
- To forecast weekly department-wise supermarket sales based on a variety of factors, such as item pricing and temperature.
- To identify which ML algorithm is best for sales forecasting.

II. Literature Survey

The authors of [8] described how they used ML algorithms to estimate sales. In this work, ML techniques such as Gradient Boosting, KNN, and Random Forest were used to forecast supermarket sales. The results showed that the MAE, 0.425, 0.428, and 0.409, respectively, and the standard deviation, 0.018, 0.0144, and 0.0144. The authors in [9] presented forecast of Walmart's gross sales. Using the Random Forest method, this paper's substantial and exclusive work on the Walmart store allowed it to forecast sales. The authors of [10] compared and worked on supervised ML methods for forecasting shop sales. Apart from supervised ML techniques random forest, SVM, KNN, linear, and tree regressions, this article emphasizes the usefulness of data mining. The RMSE values for the year 2012 are 557863.9, 95490.48, 350586.93, 518924.75, and 62871.85, respectively. The authors of [11] have reported their findings on the use of ML algorithms for autonomous product classification. This essay offers a method for classifying products according to the classification schemes of the three North American nations. The precision values reported by the authors were 0.93 for Naïve Bayes, 0.95 for Linear Regression, and 0.96 for SVM. Owing to the significance of forecasting across a wide range of sectors, a great number of diverse methodologies have been used in the past, including statistical, hybrid, and ML models. Statistical methods, like auto regressive moving average and auto regressive integrated moving average, are handy. Bipartisan visual clusters, which grouped several warehouses on sales activity, were tested by [12]. Bayesian network is used to handle the application, they were able to offer an improved forecasting experience. In [13] authors surveyed food sales forecasting using ML techniques. In this survey, they talked about data analyst design choices including output variables, temporal granularity, and input variables [4]. To improve the accuracy of demand forecasting, the authors experimented on point of sale (POS) data as internal data and even included external data including surroundings into account. For assessment, they took into account various ML techniques, including Decision Forest Regression, Bayesian Linear Regression, and Boosted Decision Tree Regression. [14] The authors' research on patrons visiting restaurants utilizing XGBoost, k-nearest neighbor, and Random Forests was fascinating. They made many input variables from restaurant attributes and selected two real-world data sets from various booking websites. XGBoost is the best model for the dataset, according to the results. The weather has an impact on regular restaurant sales, as noted by the authors in [15]. They looked at two ML algorithms: neural networks and XGBoost. The results indicated that the XGBoost algorithm is more accurate than the other algorithm, and that by accounting for weather factors, they were able to improve their model performance by 2-4 percentage points. They had taken into account a number of elements, including weather conditions, sales history, and date features, in order to increase accuracy. [16] The majority of previous studies concentrated on sales modeling and employed training data directly, not taking into account the relationship between the training and testing data. This results in numerous errors, which lower accuracy. In order to reduce computational time and obtain effective evaluation performance, recent studies have proposed clustering strategies to partition the full forecasting data into many clusters of predictable data prior to creating predictable models. [17] SVM has been specifically used in demand forecasting, and an intelligent model that uses SVMs to address problems with allocation and discovery of new models has been developed. Authors shown in [18] that SVM performed better in predicting demand on consumer goods when compared to Neural Networks. In [19] the authors have used store, promotion and competitor data for forecasting sales of a store by using ML algorithms. This paper has predicted the sales of a drug store for 6 weeks using ML algorithms gradient boosting, random forest, KNN, logistic regression and linear regression with scores 0.1077, 0.1199, 0.3058, 0.3328 and 0.3897 respectively. In [20] the researchers have worked on the analysis of supermarket data. In [21] the authors have worked on which ML algorithm would suit best for sales prediction. This paper has extensively worked on Walmart store data and used regressor algorithms like ridge, multiple linear, lasso linear and RMSE values respectively: 1.572838e+05, 1.572616e+05, 1.57283e+05. In [22] the researchers have worked on Bigmart store data and used various ML algorithms for predicting the sales. This paper has used algorithms like linear regression, polynomial regression, XGBoost and ridge regression with RMSE values 7.4631, 2.731, 7.823, 0.032 and 1.916 respectively. In [23] the author has showed results for automating product categorization using ML. The author has chose logistic regression for this purpose since this is a multiclass classification problem and obtained a precision, recall and F1 Score of 95.57%, 95.57% and 95.57% respectively. In [24] the authors worked on sales prediction using specific ML algorithms- KNN regression and linear regression with RMSE of 2546.79 and 1898.91 on test data respectively.

III. DATASET DESCRIPTION

There were 3 datasets that were present in Kaggle [25] as base datasets. The datasets have been merged with common features to create a bigger dataset. The 3 datasets are as follows:
1. stores.csv: CSV file with features- store, type and size for 45 different stores.
2. features.csv: CSV file with attributes- Store, Date, Temperature, Fuel_Price, MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5, CPI, Unemployment, IsHoliday with 8190 data tuples.
3. train.csv has attributes a. Store, b. IsHoliday c. Weekly_Sales, d. Date,e. Dept, with 1,15,064 tuples.
Taking all these into consideration, and renaming the columns, new dataset contains 11 features and 115064 tuples with the features: ItemWeight, ItemIdentifier, ItemFatContent, OutletIdentifier, OutletType, ItemVisibility, ItemType,Item MRP, Outlet EstablishmentYear, OutletSize, OutletLocationType. For the purpose of automatic product classification, flipkart dataset [26] is considered which consists of 20,000 rows and 15 columns like id, url, name, hierarchy, price, discounted price, image, description, category etc. out of which 2 features description and product category are selected.

IV. METHODOLOGY

The steps which were performed are as i. tapping the data, ii. preprocessing, iii. ML model building, iv. data visualisation and analysis, v. performance of models vi. and deploying the model.

- *Data collection*: The dataset was taken from an internet source- Kaggle.com which was taken from 45 different stores and was modified a little bit to merge. It had various attributes like product id, product type, outlet type etc.
- *Data pre-processing*: The data that was collected cannot be taken as it is, since it had few irrelevant values(negative numbers where they're not applicable), cleaning the data etc.
- *Handling missing data*: There were missing values in various features and an average of the existing values was taken to fill in that data.
- *Data reduction*: In this step, the huge dataset will be reduced into a smaller one.

- Feature extraction involves transforming new input data into essential attributes while preserving data integrity, confidentiality, and storage. It aims to maintain the original information of the dataset while processing and extracting cleaned attributes necessary for analysis at each stage.

A. ALGORITHMS

Random Forest Regression is a supervised method that employs ensemble learning for regression tasks. By combining predictions from multiple machine learning algorithms, it aims to deliver more accurate predictions compared to individual models. This technique performs effectively across various problems, especially those featuring non-linear relationships among attributes [27].
Ridge regression is a model-tuning method employed for analyzing datasets showing multicollinearity. This technique applies L2 regularization to address issues arising from multicollinearity, where predicted values may significantly deviate from real values due to impartial least-squares and high variances. The standard regression equation, $Y = XB + e$, underpins all types of regression machine learning models. Here, Y represents the dependent variable, X denotes the independent variables, B represents the regression coefficients to be determined, and e stands for the residual errors [28].
Lasso regression: Least Absolute Shrinkage and Selection Operator is what "LASSO" stands for. One method of regularization is lasso regression. For a more accurate forecast, it is preferred to regression approaches. Shrinkage is used in this model. When data values shrink toward the mean, or a central point, this is called shrinkage. Simple, sparse models—i.e., models with fewer parameters—are encouraged by the lasso process. When you wish to automate specific steps in the model selection process, such as the deletion of variable selection parameters, this specific kind of regression is ideal for models exhibiting high degrees of multicollinearity. The L1 regularization technique is used in Lasso Regression [29].
An technique called the Linear Support Vector Machine (Linear SVC) looks for a hyperplane that maximizes the distance between categorized samples. This technique works well with a high number of samples and conducts classification using a linear kernel function. By comparison, the Linear SVC model contains more parameters than the SVC model, including a loss function and penalty normalization that applies 'L1' or 'L2'. Since linear SVC is based on the kernel linear technique, the kernel method cannot be altered [30].
Naïve Bayes is widely employed for text classification tasks, leveraging high-dimensional training datasets. As one of the simplest and most efficient classification algorithms, Naïve Bayes facilitates the swift development of machine learning models with quick prediction capabilities. Operating as a probabilistic classifier, it predicts outcomes based on the likelihood of an event occurring. Common applications of the Naïve Bayes algorithm include sentiment analysis, article classification, and spam filtering [31].
Random Forest Classification operates by constructing multiple decision trees on different subsets of the provided dataset and then averaging their predictions to enhance predictive accuracy. Instead of relying on a single decision tree, random forest makes predictions based on the majority vote of projections from each tree. By aggregating the predictions of multiple trees, random forest mitigates the risk of overfitting and boosts accuracy, particularly as the number of trees in the forest increases. [32].

B. Analysis tasks performed

1. Automatic product categorisation

In this, given the description for a product, the ML model classifies it into a unique category out of the 10 categories as displayed in Fig. We have used Naïve Bayes, LinearSVC and Random Forest algorithms for this purpose- out of which LinearSVC has shown better results which is shown in Table 1.

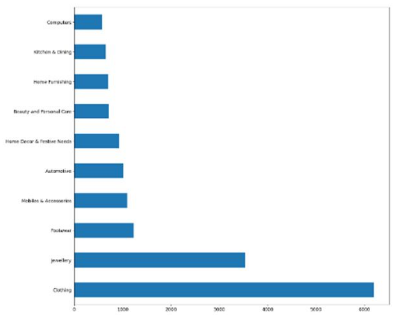


Fig 1: Categories of Products

TABLE I: ALGORITHMS USED AND RESULTS OBTAINED FOR AUTOMATIC PRODUCT CLASSIFICATION

Algorithm	WordLevel tf-idf	Bigram tf-idf	4-gram tf-idf	Charlevel tf-idf
Random forest	0.978	0.981	0.980	0.970
Naïve Bayes	0.953	0.973	0.969	0.876
Linear SVC	0.993	0.992	0.993	0.990

2. Identifying and plotting how various factors have affected the sales as shown in Fig. 2

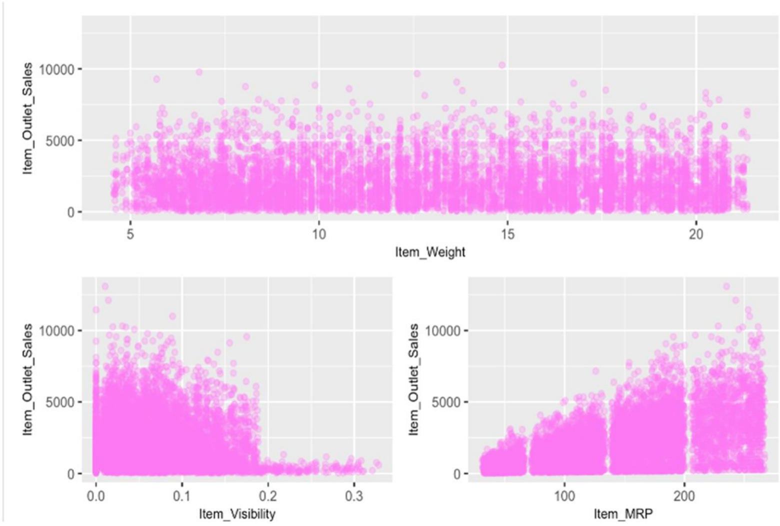


Fig 2: Plots of how various factors effecting sales

The association between item_outlet_sales and item_MRP is linearly positive. The data exhibits some non-constant variance or dispersion. Sales and item weight primarily have a non-linear relationship. However, there is some positive linear association with consumables. Sales and item visibility exhibit a somewhat positive linear connection. However, the plots highlight the data's significant non-constant volatility. In all three graphs, there are anomalies in the consumable and non-consumable item categories.

3. Plotting a correlation plot among all the attributes as shown in Fig 3.

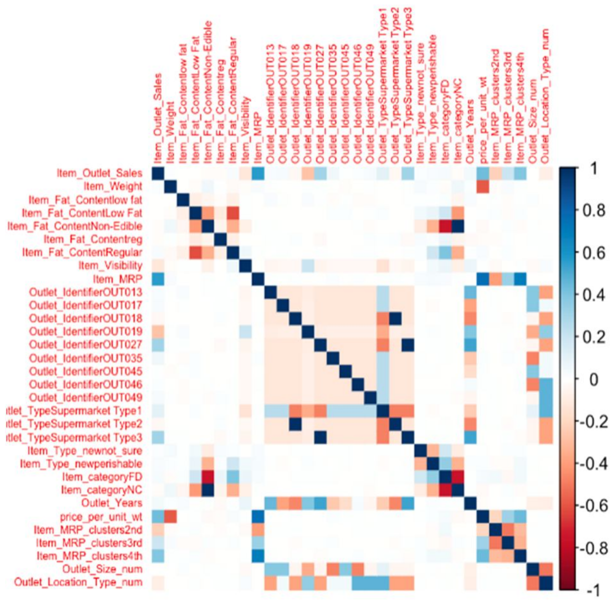
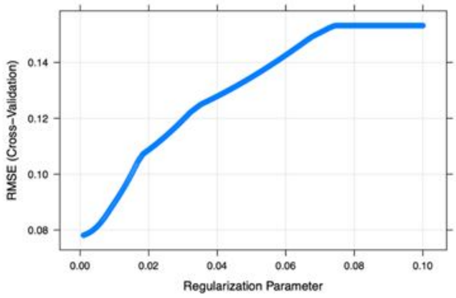
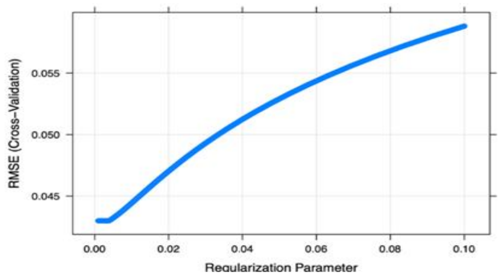


Fig 3: Correlation plot of all attributes

4. Identifying and plotting the RMSE for lasso regression, ridge regression and random forest models as shown in Fig. Ridge regression with RMSE 0.055, Random forest: 0.077, Lasso: 0.145



Lasso regression



Ridge regression

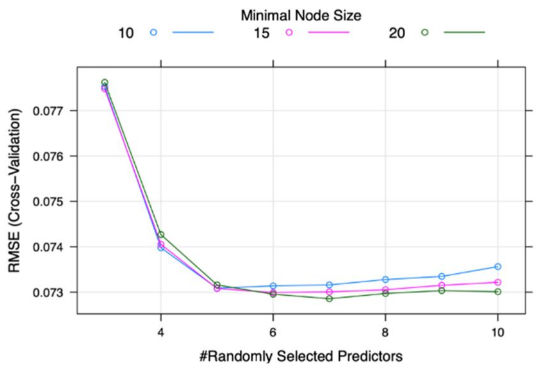


Fig 4: RMSE of Various Algorithms

V. Conclusion And Future Scope

The findings of this paper highlight the prediction of future sales of any type of store by department, weekly wise. The various models used for this task are: regressors ridge, random forest and lasso; the best being random forest. This paper also includes the task of automatic product classification- where a product will be classified into pre-defined classes(supervised learning), given the description of the product. There were 10 main categories identified like clothing, jewelry, household etc. For this purpose, LinearSVC, random forest and Naïve Bayes were considered, the best being LinearSVC.

This paper also finds out what are the factors that effects the sales mostly in each category. For example, when we look at the item fat content, the low-fat items were mostly sold when compared to regular fat items; household items were mostly sold when compared to other items like food etc. Another aspect is to look at how various factors like item price, outlet location etc. have effected the sales. For this purpose, a correlation plot was plotted and it was observed that item price has effected the sales mostly, which is pretty obvious.

The future work which can be added to this could be:

- Automating the restocking of products based on sales predicted.
- Creating a website that can be accessed by all the stores so that the customers can know about the stock, bestsellers, and offers depending on the sales predicted.

With these future works, the task would be completed in a way that can be extended for the use of everyone.

References

[1] Patrick Bajari, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. ML methods for demand estimation. American Economic Review, 105(5):481–85, 2015.

[2] Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. Mnfg & Service Operations Management, 18(1):69–88, 2016.

[3] Ankur Jain, Manghat Nitish Menon, and Saurabh Chandra. Sales forecasting for retail chains, 2015.

[4] Grigorios Tsoumakas. A survey of ML techniques for food sales prediction. Artificial Intelligence Review, 52(1):441–447, 2019

[5] Xiaogang Su, Xin Yan, Chih-Ling Tsai. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3):275–294, 2012.
[6] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. Journal of the American statistical association, 83(404):1023–1032, 1988.
[7] Zheng Li, Xianfeng Ma, and Hongliang Xin. Feature engineering of machine learning chemisorption models for catalyst design. Catalysis today, 280:232–238, 2017.
[8] Odegua R. Applied Machine Learning for Supermarket Sales Prediction. Project: Predictive Machine Learning in Industry. 2020.
[9] Mounika S, Sahithi Y, Grishmi D, Sindhu M, Ganesh P. Walmart Gross Sales Forecasting Using Machine Learning. Jl of Advanced Research in Technology and Management Sciences. 2021;3(4):22-7.
[10] Raizada S, Saini JR. Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting. International Journal of Advanced Computer Science and Applications. 2021;12(11).
[11] Roberson A. Applying Machine Learning for Automatic Product Categorization. Journal of Official Statistics. 2021 Jun 1;37(2):395-410.
[12] Grigorios Tsoumakas. A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1):441–447, 2019.
[13] İrem İşlek and Şule Gündüz Öğüdücü. A retail demand forecasting model based on data mining techniques. In 2015 IEEE 24th Intl Symposium on Industrial Electronics (ISIE), pages 55–60. IEEE, 2015.
[14] Takashi Tanizaki, Tomohiro Hoshino, Takeshi Shimmura, and Takeshi Takenaka. Demand forecasting in restaurants using machine learning and statistical analysis. Procedia CIRP, 79:679–683, 2019.
[15] Xu Ma, Yanshan Tian, Chu Luo, and Yuehui Zhang. Predicting future visitors of restaurants using big data. In 2018 Intl.conf. on Machine Learning and Cybernetics (ICMLC), vol 1, pp 269–274. IEEE, 2018.
[16] Mikael Holmberg and Pontus Halldén. Machine learning for restaurant sales forecast, 2018.
[17] I-Fei Chen and Chi-Jie Lu. Sales forecasting by combining clustering and machine-learning techniques for computer retailing. Neural Computing and Applications, 28(9):2633–2647, 2017.
[18] Malek Sarhani, Abdellatif El Afia. Intelligent system-based support vector regression for supply chain demand forecasting. In 2014 2nd World Conf. on Complex Systems (WCCS), pages 79–83. IEEE, 2014.
[19] Zhou Q, Huang K, Huang D. Forecasting sales using store, promotion, and competitor data.
[20] https://towardsdatascience.com/supermarket-data-analysis-with-r-5284be9541c4
[21] https://www.kaggle.com/code/yasserh/walmart-sales-prediction-best-ml-algorithms
[22] Ranjitha P, Spandana M. Predictive analysis for big mart sales using machine learning algorithms. In 2021 5th Intl Conf on Intelligent Computing and Control Systems (ICICCS) (pp. 1416-1421). IEEE.
[23] https://medium.com/code-contour-by-bfa/automating-product-categorization-c7ddc1561a30
[24] Kohli S, Godwin GT, Urolagin S. Sales prediction using linear & KNN regression. In Advances in Machine Learning & Computational Intelligence: Proc. of ICMLCI 2021 (pp. 321-329). Springer
[25] https://www.kaggle.com/datasets/singhbirender/storescsv, https://www.kaggle.com/datasets/berkayalan/retail-sales-data?select=store_cities.csv%2C, https://www.kaggle.com/datasets/berkayalan/retail-sales-data,
[26] https://www.kaggle.com/datasets/PromptCloudHQ/flipkart-products
[27] Segal, M. R. (2004). Machine learning benchmarks and random forest regression.
[28] McDonald, G. C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), 93-100.
[29] Ranstam, J., & Cook, J. A. (2018). LASSO regression. Journal of British Surgery, 105(10), 1348-1348.
[30] Rosipal, R., Trejo, L. J., & Matthews, B. (2003). Kernel PLS-SVC for linear and nonlinear classification. In Proc of the 20th Intl Conf on Machine Learning (ICML-03) (pp. 640-647).
[31] Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in AI (Vol. 3, No. 22, pp. 41-46).
[32] Pal, M. (2005). Random forest classifier for remote sensing classification. Intl journal of remote sensing, 26(1), 217-222.