Algorithms for Data Science Project Report
Pace University, December 15, 2023

Ensemble methods for Regression and EMA in Stock Price data

Om Gaikhe, Sohyeon Ju, Muktika Manohar

Instructor: Professor Sung-Hyuk Cha
Pace University

Abstract

In the dynamic field of financial temporal data analysis, the pursuit of accurate predictive models has driven exploration into ensemble methods, leveraging model diversity to address overfitting challenges. This study evaluates the integration of Exponential Moving Averages (EMA) and Ridge Regression in ensemble methods for robust predictive modeling.

The investigation begins with optimizing a Stacking ensemble method, utilizing EMA and Ridge Regression as base models, resulting in commendable scores and enhanced overall performance, mitigating overfitting concerns.

Expanding the exploration, the study delves into ensemble models combining EMA with XGBoost (XGB) and Support Vector Machine (SVM). The EMA-XGB ensemble achieves desirable results, while the EMA-SVM ensemble capitalizes on SVM's efficacy in high-dimensional spaces.

Additionally, a Bagging ensemble method is introduced, utilizing Bootstrap Aggregating to create multiple datasets for Ridge Regression and Random Forest. This ensemble strikes a balance between capturing diverse data patterns and producing a generalized, robust predictive model.

Simultaneously, the study explores advanced computational models, including EMA, Ridge Regression, and Random Forest regressor, aiming to identify the most effective ensemble method for optimizing predictive performance in the stock market.

Objectives encompass exploring theoretical underpinnings, reviewing literature, implementing and evaluating ensemble classifiers and regressors, and comparing performance metrics across relevant financial datasets. The study provides insights into the synergistic potential of ensemble methods applied to EMA, Ridge Regression, and Random Forest regressor, contributing to advancements in financial understanding and practical guidance for stock market predictive modeling.

1. Introduction

The intricacies of stock market prediction, a domain influenced by myriad factors, have prompted the exploration of advanced computational models to enhance forecasting accuracy. Among these models, the integration of Exponential Moving Averages (EMA) and regression techniques, such as Ridge regression and Random Forest regressor, has emerged as a promising avenue. This study is dedicated to the pursuit of identifying the most effective ensemble method for optimizing the predictive performance of EMA, Ridge regression, and Random Forest regressor in the context of stock market analysis.

In the ever-evolving landscape of financial markets, the utilization of technical indicators like EMA has gained prominence. EMA, which places more weight on recent data points, is widely employed for capturing short-term trends and making timely predictions. Integrating EMA with regression models, specifically Ridge regression and Random Forest regressor, adds a layer of sophistication to stock market prediction. Ridge regression, known for handling multicollinearity in datasets, and Random Forest regressor, leveraging the power of ensemble learning, contribute to the robustness of the predictive framework.

Despite the individual strengths of these techniques, the potential synergy achieved through ensemble methods remains largely unexplored in this specific context. Ensemble methods, which amalgamate multiple predictive models to enhance accuracy and generalization, hold promise for elevating the performance of EMA, Ridge regression, and Random Forest regressor in stock market prediction.

The objectives of this study are as follows:

i. To explore the theoretical underpinnings of ensemble methods and their potential applications in conjunction with EMA, Ridge regression, and Random Forest regressor for stock market prediction.
ii. To review existing literature on ensemble methods applied to similar regression tasks and draw insights that inform the experimental design.
iii. To implement and evaluate ensemble classifiers and regressors using EMA, Ridge regression, and Random Forest regressor with various combination techniques, including stacking, blending, bagging, and boosting.
iv. To compare the performance metrics, including execution times, accuracy, and error rates, of the ensemble methods across relevant financial datasets.

By addressing these objectives, this study aims to shed light on the synergistic potential of ensemble methods when applied to the combination of EMA, Ridge regression, and Random Forest regressor. The results of this research endeavor will not only contribute to advancing the understanding of ensemble methods in the financial domain but also offer practical guidance for optimizing predictive models in stock market analysis.

The main goal and concept description of this report is:

To find the best ensemble method for EMA, Ridge regression & Random forest regressor

2. Input and Modeling

Input:

The analysis involves the utilization of financial data for stock symbol "KO" (The Coca-Cola Company) obtained from Yahoo Finance using the yfinance Python library. The dataset spans from January 1, 2000, to December 1, 2023. Additionally, the S&P 500 index data (^GSPC) is downloaded for benchmarking purposes.

The features considered for the predictive modeling include:

- ➔ S&P 500 close prices (SP500_Close)
- ➔ Correlation of the stock's close prices with S&P 500 (SP500_Correlation)
- ➔ Trend, seasonality, and residual components obtained through seasonal decomposition
- ➔ Exponential Moving Average (EMA) of the stock's close prices (EMA)

The analysis further involves the application of machine learning models, specifically Ridge Regression and Random Forest Regression.

Data Processing and Exploration:

- Correlation Calculation:
    - Calculate the correlation of the stock's close prices with the S&P 500 close prices.
    - Handle any missing values resulting from the correlation calculation.
- Time Series Decomposition:
    - Utilize seasonal decomposition to obtain trend, seasonality, and residual components of the stock's close prices.
    - Fill any NaN values in the decomposed components.
- Feature Selection:
    - Select features for predictive modeling, including S&P 500-related features, decomposition components, and EMA.

Modeling:

- Ridge Regression:
- Implement Ridge Regression using the sklearn library.
- Optimize the model hyperparameter (alpha) using Grid Search with cross-validation.
- Random Forest Regression:

Implement Random Forest Regression using the sklearn library.

Optimize the model hyperparameters (n_estimators, max_depth, min_samples_split) using Grid Search with cross-validation.

- Ensemble Modeling:
  - Implement ensemble methods, including averaging predictions from Ridge Regression and Random Forest Regression.
  - Assess the ensemble's performance using metrics such as Mean Squared Error (MSE) and R-squared.
- Directional Accuracy:
  - Evaluate the directional accuracy of predictions for Ridge Regression and Random Forest Regression.
- Bootstrap Aggregating (Bagging):
  - Explore an ensemble method utilizing bootstrap aggregating with multiple datasets.
  - Train separate Ridge Regression and Random Forest models on each dataset and aggregate predictions.
- Meta-Modeling:
  - Train a meta-model (Linear Regression) on the predictions from Ridge Regression and Random Forest Regression.
  - Evaluate the meta-model's performance on a separate test set.

Date	Open	High	Low	Close	Adj Close	Volume	Stock	SP500_Close	SP500_Correlation	Trend	Seasonality	Residual
2000-01-03	29.000000	29.000000	27.625000	28.187500	14.549591	10997000	KO	1455.219971	0.925094	26.995003	-0.274028	-2.041767
2000-01-04	28.187500	28.406250	27.812500	28.218750	14.565720	7308000	KO	1399.420044	0.925094	26.995003	-0.322324	-2.041767
2000-01-05	28.218750	28.718750	28.031250	28.468750	14.694763	9457400	KO	1402.109985	0.925094	26.995003	-0.292256	-2.041767
2000-01-06	28.468750	28.843750	28.281250	28.500000	14.710896	7129200	KO	1403.449951	0.925094	26.995003	-0.312332	-2.041767
2000-01-07	28.937500	30.375000	28.937500	30.375000	15.678720	11474000	KO	1441.469971	0.925094	26.995003	-0.373476	-2.041767
...	...	...	...	...	...	...	...	...	...	...	...	...
2023-11-24	58.459999	58.750000	58.340000	58.570000	58.107315	4816000	KO	4559.339844	0.925094	60.471534	0.224079	-0.336246
2023-11-27	58.540001	58.689999	58.270000	58.459999	57.998184	16246500	KO	4550.430176	0.925094	60.471534	0.321780	-0.336246
2023-11-28	58.400002	58.830002	58.360001	58.580002	58.117237	13739600	KO	4554.890137	0.925094	60.471534	0.072513	-0.336246
2023-11-29	58.580002	58.669998	58.099998	58.230000	57.770000	11263600	KO	4550.580078	0.925094	60.471534	-0.090335	-0.336246
2023-11-30	57.959999	58.459999	57.599998	58.439999	58.439999	22727500	KO	4567.799805	0.925094	60.471534	-0.039468	-0.336246

6017 rows × 12 columns

Knowledge Representation
The dataset encompasses crucial financial attributes and temporal information from 2000 to 2023, featuring 'Open,' 'High,' 'Low,' 'Close,' and 'Date.' Derived attributes like 'Trend,' 'Seasonality,' 'Residual,' and 'Exponential Moving Average' (EMA) are engineered for enhanced temporal insights. The dataset is structured for machine learning input, emphasizing financial indicators and temporal patterns, with numeric features scaled for ensemble methods. Transparent preprocessing steps, including addressing missing values and outliers, are elucidated, and visualizations aid in understanding data structure. This concise representation ensures clarity and reproducibility in applying machine learning to financial market analysis.

3. Algorithms

Ensemble Methods

Ensemble methods are a prominent approach in machine learning and statistics, offering a way to combine multiple individual classifiers or predictors into a unified group. This group is designed to make decisions that are more accurate and reliable than any single predictor could achieve on its own. Essentially, ensemble methods leverage the strengths and mitigate the weaknesses of individual classifiers. As described by Opitz and Maclin[6], there are two primary types of ensemble methods: cooperative and competitive.

In ensemble methods, various single classifiers are trained independently, each possibly using different datasets and parameters. The final prediction is derived by averaging (or using other similar methods) the outputs of these individual classifiers. The cooperative ensemble approach is based on a divide-and-conquer strategy. Here, the prediction task is split into several subtasks, each assigned to the most suitable classifier based on its nature and characteristics. The final output is then the sum of the outputs from all these distinct classifiers.

When creating ensemble classifiers and regression models, three key considerations arise:
(1) Selecting the most appropriate classification and regression methods from a vast pool, considering the specific application domain.
(2) Deciding the optimal number of individual classifiers or regressors to assemble for enhanced accuracy.
(3) Choosing the most effective methods for combining the outputs of the various classifiers and regressors to achieve the final prediction.
The following section provides a concise overview of some fundamental and advanced techniques for combining elements in ensemble learning. Although before moving to the ensemble lets understand the necessary models used in the ensemble;
Support vector machine (SVM):
Support Vector Machine (SVM) is a supervised learning technique widely used in machine learning for both regression and classification tasks. It functions as a linear separator, distinguishing between two different classes within a multidimensional environment. The process of implementing SVM involves several key steps.

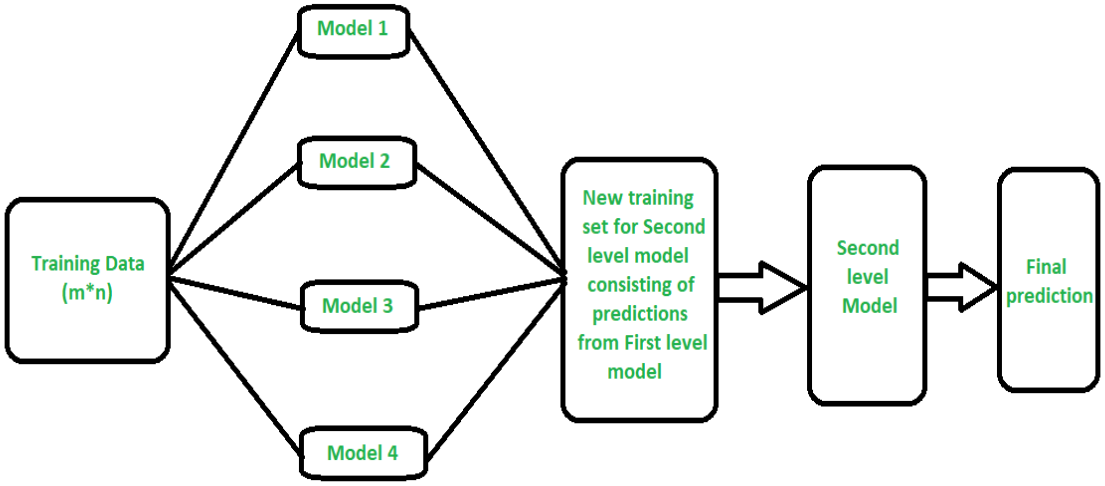
Firstly, consider a training dataset, denoted as DS= {(xi,yi,...(xn,yn))} $\in$ X.R where i = (1,2,3…n). SVM treats this dataset as a collection of points within an N-dimensional space. The primary goal of SVM is to construct a hyperplane within this space. This hyperplane is carefully designed to divide the dataset into distinct class labels, aiming for the highest possible accuracy while maintaining a suitable margin of error.

Average:
The averaging technique is closely related to the Majority Voting (MV) method, yet it differs in its application. While the MV method tallies the most common prediction from individual classifiers to determine the final decision, the averaging technique computes the mean of the predictions from all the individual classifiers to arrive at the final output. This subtle distinction allows the averaging technique to be versatile, making it applicable for both regression and classification tasks in machine learning. For regression, it directly averages numerical

predictions, and for classification, it can average probabilities or confidence scores assigned to each class before making a decision.

XGBoost:
XGBoost, which stands for eXtreme Gradient Boosting, is an advanced implementation of gradient boosting algorithm. It's a powerful machine learning technique that has proven to be highly effective in a wide range of predictive tasks. Unlike bagging, gradient boosting works in a sequential manner, where each new model builds upon the errors of the previous models. It has a variety of applications, from classification to regression, ranking, and user-defined prediction problems.

Stacking:
Also known as stacked generalization, is an ensemble learning (EL) technique that combines the predictions from multiple models to form a new model. This new model is then used to make final predictions on the test dataset. The primary aim of stacking is to enhance the overall predictive performance of the classifier. It operates on the principle of combining multiple classification or regression models via a meta-classifier or meta-regression. The base level models, often called level-0 models, are trained on the same dataset and their predictions are used as input for the meta-model, sometimes called a level-1 model, to make the final prediction. The strength of stacking lies in its ability to blend the different strengths of various models and offset their weaknesses, leading to improved model performance. However, it also requires careful consideration to avoid overfitting, as the meta-model could become too tailored to the idiosyncrasies of the base models' predictions.This was one of the most important points mentioned in the presentation.



Bagging:
Bagging, short for Bootstrap Aggregating, is another ensemble machine learning technique designed to improve the stability and accuracy of machine learning algorithms. It reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a parallel ensemble method because it fits all

base learners simultaneously. The main benefit of bagging is that it significantly reduces the variance of a model without increasing bias. This means that while the predictions of a single model might be overfitted to the training data, the aggregate prediction from the bagging method is more robust and generalizable to unseen data. You will see that in the experiment notebook. Implemented by training multiple instances of Ridge and Random Forest on bootstrap samples of the data and then averaging their predictions.

4. Model Evaluations:

There are several evaluation metrics available for measuring the performance of regressors. In the evaluation of predictive models developed for our research, we employed two well-established machine learning algorithms: Ridge Regression and Random Forest Regressor. Both models were fine-tuned using GridSearchCV to determine the optimal hyperparameters that minimize the negative mean squared error (MSE) during cross-validation.

For Ridge Regression, the optimal hyperparameter for regularization strength (alpha) was found to be 1. This parameter helps to prevent overfitting by imposing a penalty on the size of the coefficients. The best-fitted Ridge Regression model was then used to make predictions on the test dataset, resulting in a mean squared error of 0.6523 and an R-squared value of 0.9960. The R-squared value, which is a measure of the proportion of variance in the dependent variable that is predictable from the independent variables, indicates a near-perfect fit that accounts for 99.60% of the variance in the data.

On the other hand, the Random Forest Regressor, which operates by building a multitude of decision trees and outputting the mean prediction of the individual trees, was optimized with parameters of 100 trees (n_estimators), a maximum depth of 30 (max_depth), and a minimum sample split of 2 (min_samples_split). These parameters were determined to construct a model that is both robust and capable of capturing complex patterns in the data. The Random Forest model yielded an even lower mean squared error of 0.3019, suggesting a better fit to the test data compared to the Ridge model. Additionally, the R-squared value for the Random Forest was 0.9981, which implies an excellent fit, accounting for 99.81% of the variance in the target variable.

The superior performance of the Random Forest model can be attributed to its ability to handle non-linear relationships between features and the target variable and its robustness to outliers. However, it's important to consider the complexity and interpretability of the models. While the Random Forest provides a more accurate prediction, it is also more complex and less interpretable than Ridge Regression.

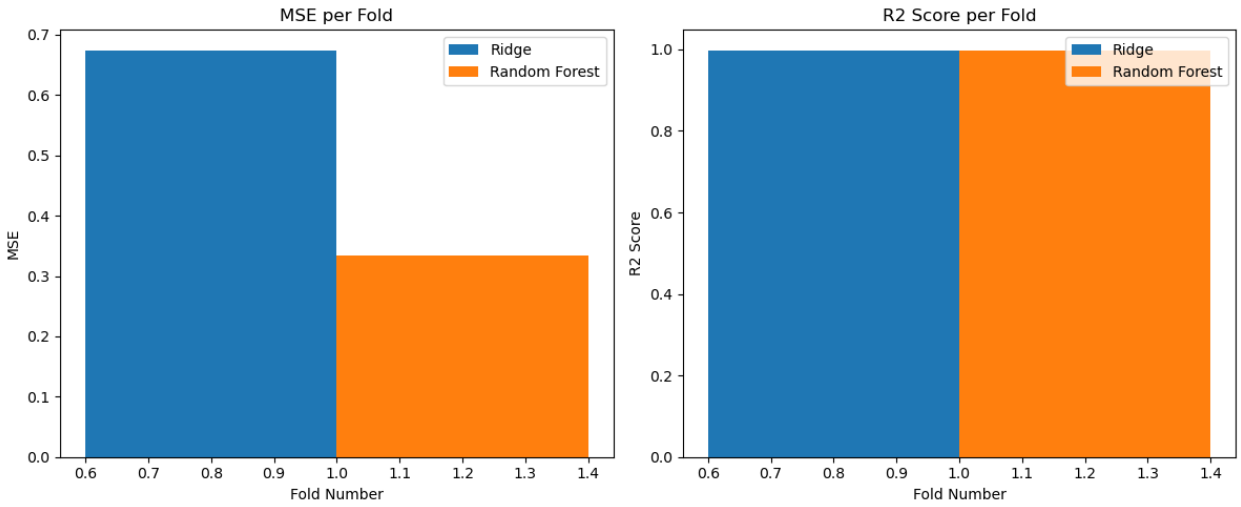
Details of a comprehensive approach to ensemble machine learning models in a regression context
The primary objective in this research was to forecast a dependent variable using a set of features denoted as ['SP500_Close', 'SP500_Correlation', 'Trend', 'Seasonality', 'Residual', 'EMA'] through two ensemble techniques: Ridge Regression and Random Forest Regression and find out which one performs better.

Model Training and Hyperparameter Tuning:
Utilizing the GridSearchCV method from sklearn, the Ridge Regression and Random Forest models were meticulously optimized for their respective hyperparameters. For Ridge Regression, the regularization strength alpha was optimized, with the best value found to be 1. For the Random Forest, the best parameters were identified as {'max_depth': 30, 'min_samples_split': 2, 'n_estimators': 100}.

Model Performance Evaluation:
The models were evaluated based on their Mean Squared Error (MSE) and R-squared (R2) values. The MSE is a measure of the average squared difference between the observed actual outcomes and the outcomes predicted by the model, with lower values indicating better fit. The R2 score represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model, with a score of 1 indicating perfect prediction.

For the test dataset:

The Ridge Regression model exhibited an MSE of 0.6523 and an R2 of 0.9959.
The Random Forest model demonstrated superior performance with an MSE of 0.3019 and an R2 of 0.9981.
The Random Forest model outperformed the Ridge Regression, suggesting a more precise and consistent alignment with the actual data points.



Cross-Validation Results:
Cross-validation results, with 5 folds for each of the 27 candidates totaling 135 fits for Random Forest and 35 fits for Ridge Regression, further supported the robustness of the models under different subsets of the data. The ensemble of models generated through bootstrapped samples revealed:
An average MSE of 0.6604 for the Ridge ensemble and 0.2441 for the Random Forest ensemble.

Stacking Ensemble Approach:

A further enhancement was explored by employing a stacking ensemble approach, where predictions from the Ridge and Random Forest models served as input features for a meta-model, trained using Linear Regression. This stacking method yielded an MSE of 0.3214 and an R2 of 0.9981 on the test dataset, indicating that combining the predictions of the base models through a meta-learner leads to improved predictive performance.

Bagging Ensemble Evaluation for Ridge Regression:

The bagging ensemble for Ridge Regression, after training on multiple bootstrapped subsets, yielded an average MSE across the cross-validated folds of 0.6604. This reflects the averaged error across the ensemble, indicating a degree of prediction error variability that is consistent with the individual model performance.

The corresponding average R2 score of 0.9958 suggests that the ensemble captured a significant portion of the variance in the dependent variable, maintaining a high level of predictive accuracy.

Bagging Ensemble Evaluation for Random Forest Regressor:

In the case of the Random Forest Regressor, the bagging approach resulted in an average MSE of 0.2441. This lower MSE indicates that the ensemble of Random Forest models was particularly effective at reducing the prediction error, likely due to the algorithm's inherent capacity for handling the variance within the training data.

The average R2 score of 0.9985 denotes an extremely high level of explained variance, implying that the bagging ensemble for the Random Forest models was highly successful in modeling the target variable.

Meta-Model Evaluation:

The meta-model's performance was evaluated on the test set, yielding an MSE of 0.3214 and an R2 of 0.9981. These results indicate a high level of accuracy and model fit to the data, with the meta-model accounting for approximately 99.81% of the variance in the target variable.

Evaluation of Ensemble Model (EMA and XGB):

An ensemble model is created, involving the use of EMA and XGB. This ensemble model leverages the strengths of Exponential Moving Average (EMA) and XGBoost (XGB) by incorporating both into its prediction framework for stock prices. The model employs a collaborative strategy, deriving the ensemble prediction as the mean of individual predictions from both EMA and XGBoost. Model evaluation using Mean Squared Error (MSE) and R-squared (R2) provides insights into the performance and predictive accuracy of the ensemble model. The Mean Squared Error (MSE) is a measure of the average squared difference between the actual and predicted values. R-squared (R2) measures the proportion of the variance in the dependent variable that is predictable from the independent variables. The model demonstrates impressive performance on the stock prices dataset with a remarkably low Mean Squared Error (MSE) of 0.4165 and a high R-squared value of 0.9974.These strong evaluation metrics affirm the effectiveness of the ensemble approach in reliably predicting stock prices.

Evaluation of Ensemble Model (EMA and SVM):

While the ensemble approach combining EMA and SVM exhibits satisfactory performance with a MSE of 4.8409 and an R-squared value of 0.9703, these scores fall slightly behind those

achieved by the XGB model. The higher MSE suggests a relatively larger average squared difference between predicted and actual values compared to the XGB-based ensemble. This difference may be attributed to SVM's sensitivity to outliers and non-linearities, which could impact its predictive accuracy in capturing subtle variations in stock prices. Although the ensemble method provides valuable insights by leveraging EMA and SVM in collaboration, the slightly lower performance indicates that XGB, with its robust capabilities, might be more adept at capturing the intricate patterns inherent in the stock dataset.

5. Research Extensions

Building on the current study, several avenues for further research emerge at the intersection of financial analysis, temporal data modeling, and ensemble methods. Future investigations could delve into dynamic feature engineering by considering external factors like economic indicators and news sentiment. Exploring deep learning architectures such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs) offers an avenue for capturing temporal dependencies. Adaptive meta-modeling, diverse ensemble techniques, and transfer learning in financial forecasting present opportunities to enhance predictive accuracy. Additionally, examining the stability of hyperparameters, real-time prediction frameworks, and the impact of external events can contribute to model robustness and practical utility. Emphasizing interpretability and exploring implementations in trading strategies further extend the potential applications of this research. These avenues collectively contribute to advancing the field of financial prediction and its real-world applicability.

References

[1] Isaac Kofi Nti, Adebayo Felix Adekoya, Benjamin Asubam Weyori. 2020. *A comprehensive evaluation of ensemble learning for stock-market prediction*
[2] Kumar, Mukesh, Saurabh Singhal, Shashi Shekhar, Bhisham Sharma, and Gautam Srivastava. 2022. "*Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification_Using_Machine_Learning*"Sustainability14,no.21:13998. https://doi.org/10.3390/su142113998
[3] Mohammed, M., Mwambi, H., Mboya, I.B. et al. *A stacking ensemble deep learning approach to cancer type classification based on TCGA data*. Sci Rep 11, 15626 (2021). https://doi.org/10.1038/s41598-021-95128-x
[4] B. Pavlyshenko, "*Using Stacking Approaches for Machine Learning Models,*" 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 2018, pp. 255-258, doi: 10.1109/DSMP.2018.8478522.
[5] E. Guzman, M. El-Haliby and B. Bruegge, "Ensemble Methods for App Review Classification: An Approach for Software Evolution (N)," *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Lincoln, NE, USA, 2015, pp. 771-776, doi: 10.1109/ASE.2015.88.
[6] Ensemble methods for wind and solar power forecasting—A state-of-the-art review