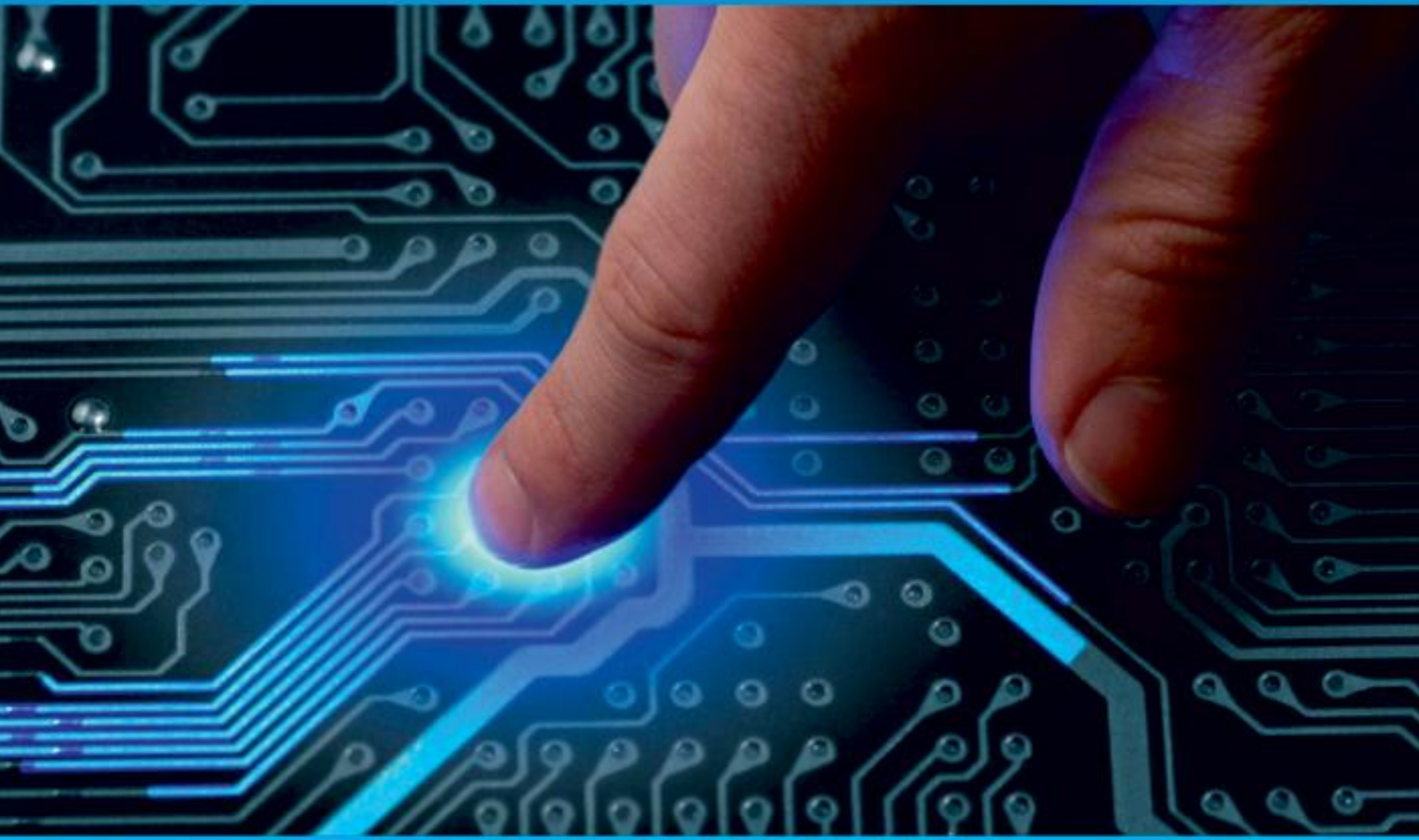
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Special Issue 1, March 2024

1st International Conference on Machine Learning, Optimization and Data Science

Organized by

Department of Computer Science and Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, India

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

Predicting Pizza Costs: An Evaluation of Random Forest and TPOT AutoML

Abhishek Singh[1], Zohaib Hasan [2], Nirdesh Jain[3]

Associate Professor, Dept. of CSE., BGIEM, Jabalpur, India[1]

Associate Professor, Dept. of CSE., BGIEM, Jabalpur, India[2]

Associate Professor, Dept. of ECE., BGIEM, Jabalpur, India[3]

ABSTRACT: In recent years, pizza has become more and more popular in India. Pizza has been available in the US for a while, but in recent years, its craze has really taken off. Important pizza-related businesses are thus seeing great investment and development opportunities in this area. Pizza is quickly rising in popularity as the preferred food for many Indian residents, and the number of restaurants serving it is expanding. In order to anticipate pizza prices, this study assesses how effective the two machine learning techniques Random Forest and TPOT (Tree-based Pipeline Optimization Tool) AutoML are. For model training and assessment, a dataset comprising several pizza variables, like diameter, toppings, and extra ingredients, was utilized. Metrics like mean squared error and R-squared were used to evaluate the prediction accuracy of both models after they had been trained and tested according to normal protocols. The findings suggest that while TPOT AutoML performs somewhat better in some circumstances, Random Forest and TPOT AutoML both exhibit encouraging performance in forecasting pizza costs. These results demonstrate how well machine learning methods work to forecast intricate pricing schemes in the food sector.

KEYWORDS: Machine learning, Random Forest, TPOT AutoML, Data Augmentation, Food industry pricing

I. INTRODUCTION

Pizza is an Italian baked pie that is flat and has an open face. It is made of a thin crust of wheat or bread dough, seasoned tomato sauce, and cheese. Anchovies, olives, mushrooms, and other garnishes are frequently added to pizza. Based on the most recent data, the size of the worldwide pizza market was estimated to be USD 178533.6 million in 2022. Over the course of the forecast period, the market is anticipated to grow at a CAGR of 6.53%, reaching USD 260982.48 million by 2028 **[1]**.
Pizza has long been accessible in India, but it has only just emerged as one of the country's most well-liked quick snacks. The Indian pizza market has expanded at a consumer annual growth rate of 26% over the last five years **[2]**.

Even though pizza seems straightforward, figuring out how much it will cost depends on a number of factors, including size, dough type, topping selection, and extras like extra cheese or sauce. Conventional pricing techniques are prone to biases and mistakes because they frequently rely on subjective judgment and manual analysis. On the other hand, machine learning algorithms, by utilizing massive datasets and sophisticated modeling approaches, offer a chance to automate the pricing process and enhance accuracy.

The use of machine learning techniques has expanded across a range of businesses in the current era of data-driven decision-making. Among them, the food industry stands out as one where predictive modeling may provide insightful information, especially when it comes to predicting product pricing. With so many variables affecting the pricing of this well-liked food item, the prediction of pizza prices is an interesting case study.

Several algorithms may be used to learn a regression model. In this study, we examine and evaluate the efficacy of two particular techniques, the XGBoost Regressor **[3]** and the RandomForest Regressor **[4]**. Because they demand a significant amount of labeled data for training and more time and processing capacity, more sophisticated machine learning techniques like neural networks were not investigated. In the food sector, data labeling is costly and time-consuming. More pragmatic and intelligible techniques, including random forest and XGBoost, were favored. In this work, we will find out how Random Forest regression and TPOT AutoML, the two machine learning techniques, are effective for pizza price prediction. While Random Forest builds many decision trees in order to get reliable predictions, TPOT AutoML uses automatic machine learning methods to find the optimal pipeline setup, which may provide better results with less manual work.

Why Random Forest ?

For a number of reasons, Random Forest is a good choice for regression tasks like price prediction of pizza.
Robustness against Overfitting: Random Forests can handle noisy or complicated datasets without compromising performance since they are less likely to overfit than certain other models.
Non-linear references: There may be non-linear links between pizza costs and size, toppings, and extra ingredients, among other things. These non-linear correlations may be efficiently captured by Random Forests, enabling more precise predictions.
Feature value: Random Forests offer a way to quantify the value of features, which might be useful in figuring out what elements have the biggest impact on pizza pricing. This information is helpful for companies looking to maximize their pricing tactics.
Ensemble Learning: By merging many decision trees, Random Forests take use of ensemble learning to increase resilience and generalization. By using an ensemble technique, the impacts of individual tree bias and variation are lessened.
Simple to Tune: A range of hyperparameters, including the number of trees (n_estimators) and the maximum depth of trees (max_depth), may be changed in Random Forests to improve performance. This adaptability makes it possible to adjust for optimal prediction accuracy.
Handles Both Numerical and Categorical Data: Random Forests are appropriate for datasets with a variety of variable kinds, as those encountered in pizza pricing prediction tasks, since they can manage a combination of numerical and categorical information.

Overall, Random Forests are a good option for predicting pizza pricing due to their adaptability, resilience, and efficacy.

Why TPOT ?

Genetic algorithms are used by the automated machine learning (AutoML) tool TPOT (Tree-based Pipeline Optimization Tool) to find the optimal machine learning pipeline for a given dataset. For estimating pizza pricing, TPOT might be a good option for the following reasons:
Automation and Pipeline Optimization: Model selection, feature engineering, feature selection, and hyperparameter optimization are all automated using TPOT. When compared to manual testing, this approach saves time and effort, particularly for complicated datasets like the one used to predict pizza costs.
Flexibility: Regression, classification, and time series forecasting are just a few of the machine learning tasks that TPOT is capable of handling. Because of its adaptability, it may be made to meet the unique needs of predicting pizza pricing, which might entail complex correlations between different pizza qualities and prices.
Ensemble Learning: Using the predictions of several different individual models, TPOT may build ensemble models. When compared to single models, ensemble learning frequently produces better performance, which is advantageous for precisely predicting pizza costs.
Automated Data Preprocessing: TPOT manages data preprocessing tasks automatically, including feature scaling, missing value imputation, and categorical variable encoding. By doing this, the user's workload is lessened and the pipeline's robustness and efficiency are guaranteed.
Capacity to Handle complicated Datasets: Estimating pizza costs may require managing high-dimensional, complicated datasets with a variety of characteristics. More precise price forecasts may result from TPOT's capacity to handle these kinds of datasets and find pertinent characteristics.
Investigating Various Models: To determine which machine learning model performs the best given the data, TPOT investigates a large variety of machine learning techniques and setups. The accuracy of the forecast might be increased if non-obvious patterns or links in the pizza pricing dataset are found during this investigation.

In conclusion, by automating the machine learning pipeline optimization process and examining a variety of models and configurations to find the best fit for the provided dataset, TPOT provides an efficient and effective method for forecasting pizza pricing.

We compare these two algorithms' prediction powers in an effort to answer the following research questions:

Firstly, How accurate are TPOT AutoML and Random Forest regression in predicting pizza prices?
Secondly, What are the advantages and disadvantages of each strategy for encapsulating the complexity of pizza price?
Thirdly, Can machine learning algorithms manage the wide variety of factors that affect pizza prices?

This study intends to give insights into the application of machine learning approaches for pricing in the food business through comparative analysis and empirical assessment. We hope to add to the expanding corpus of research on machine learning applications in pricing analytics by illuminating how well Random Forest and TPOT AutoML anticipate pizza prices.

II. RELATED WORK

An overview of machine learning's application in the food business and the status of spoiling in ready-to-eat pizza is given in this chapter. With an emphasis on the application of machine learning to improve spoilage detection and prevention in pizza and other food items, the most recent scientific discoveries and technology advancements are discussed. When it comes to ready-to-eat pizza, the majority of food technology research focuses on how food deteriorates and if it is still edible. Da-Wen Sun, Tadhg Brosnan **[5]** deals with the Pizza quality evaluation using computer vision—Part 2 Pizza topping analysis. D. Lee, J. Kim **[6]** in their research work forecasts the Pizza sales using big data analysis. Paul Wunderlich, et.al **[7]** did a remarkable work in Enhancing Shelf Life Prediction of Fresh Pizza with Regression Models and Low Cost Sensors. Additionally, some academics have looked at using machine learning in conjunction with other technologies to assess the quality of food. For example, Darwish et al. **[8]** suggested a unique technique to accurately categorize food goods as infected or uncontaminated by merging microwave (MW) detection technology with machine learning (ML) tools like MLP and SVM. In order to assess the microbiological quality of chicken burgers, Fengou et al. **[9]** looked into the use of FTIR spectroscopy and multispectral photography in conjunction with ML algorithms. A thorough summary of current advancements in hyperspectral imaging systems for identifying sensory attributes in different foods, such as color, flaws, texture, taste, freshness, and maturity, is provided by Özdoğan et al. **[10]**.

In conclusion, the food business has found machine learning to be a helpful tool for rapidly and accurately determining the cost, quality, and safety of food. To detect defects, classify food products, and predict chemical and sensory properties, various sensors—such as microwave sensing, spectroscopy, and hyperspectral imaging—have been combined with various machine learning techniques, including support vector machines, K-nearest neighbors, deep learning, and neural networks. These techniques have shown promising results in identifying pollutants and measuring ripeness when applied to a range of food commodities, such as fruits, vegetables, meat, and dairy products
.

III. PROPOSED ALGORITHM

Random Forest

One specific Random Forest algorithm implementation utilized for regression problems is the Random Forest Regressor **[4]**. Random Forest uses many decision trees ensemble learning approach to provide predictions. Every decision tree is trained by subsection of training data which is selected randomly, and, by taking the average of the predictions made by this ensemble the final prediction is calculated. This method makes it easier to reduce overfitting and improve the overall accuracy of the model.

During training, Random Forest generates several decision trees and produces the mean forecast of each tree for regression problems. Random Forest is an ensemble learning technique. The approach is widely used for predictive modeling jobs because of its capacity to manage huge datasets with high dimensionality and prevent overfitting.

A prominent Python machine learning tool, the scikit-learn module, was used in our Random Forest regression implementation. Grid search cross-validation was used to optimize the algorithm's hyperparameters, which included the number of trees in the forest (n_estimators), the maximum depth of the trees (max_depth), and the lowest number of samples needed to divide a node (min_samples_split).

Tree-based Pipeline Optimization Tool

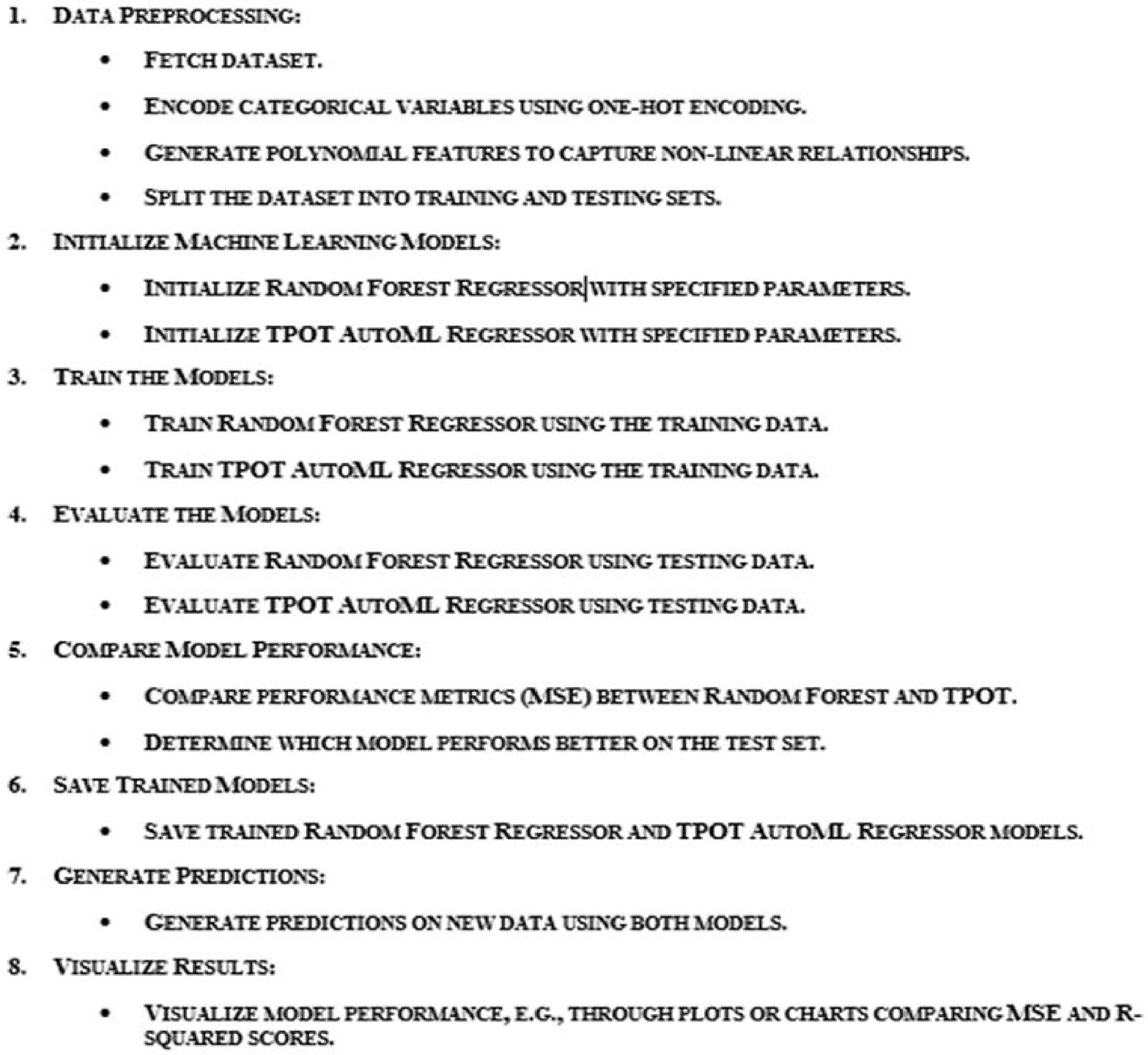
Tree-based Pipeline Optimization Tool, or TPOT Using a given dataset and prediction goal, AutoML is an automated machine learning program that finds the optimal pipeline configuration. TPOT uses an ensemble of machine learning pipelines to evolve over numerous generations using genetic programming in an effort to find the best possible mix of feature engineering methods, model algorithms, and preprocessing stages.

We investigated a variety of possible machine learning pipelines for pizza pricing prediction in our study using TPOT AutoML. We set up TPOT to operate for a predetermined maximum training duration, population size, and number of generations. The algorithm's effectiveness was assessed by looking at how well it minimized the training data's mean squared error (MSE).

IV. PSEUDO CODE

This pseudo code includes brief passages of Python code to demonstrate each phase of the process in an organized manner. It provides a concise synopsis of the tasks completed at every machine learning pipeline level. The data collection has been from Kaggle **[11]** followed by data augmentation because the data was less in number, the number of trees in the ensemble (n_estimators), the least number of samples needed to divide an internal node (min_samples_split), and the maximum depth of the trees (max_depth) are the inputs for the pseudocode method. These variables are crucial for managing the model's performance and complexity. The algorithm's initial step is to take the features (X) and labels (Y) out of the data set. After that, it divides the data in half, 70 to 30, into training and test sets.

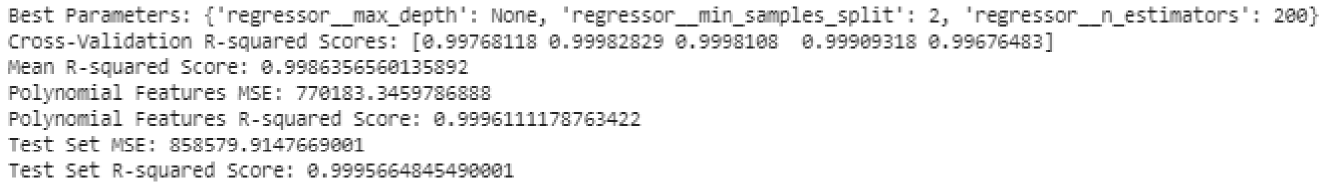
he proposed algorithm consists of the following steps:



```
1. Data Preprocessing:
    • Fetch dataset.
    • Encode categorical variables using one-hot encoding.
    • Generate polynomial features to capture non-linear relationships.
    • Split the dataset into training and testing sets.
2. Initialize Machine Learning Models:
    • Initialize Random Forest Regressor with specified parameters.
    • Initialize TPOT AutoML Regressor with specified parameters.
3. Train the Models:
    • Train Random Forest Regressor using the training data.
    • Train TPOT AutoML Regressor using the training data.
4. Evaluate the Models:
    • Evaluate Random Forest Regressor using testing data.
    • Evaluate TPOT AutoML Regressor using testing data.
5. Compare Model Performance:
    • Compare performance metrics (MSE) between Random Forest and TPOT.
    • Determine which model performs better on the test set.
6. Save Trained Models:
    • Save trained Random Forest Regressor and TPOT AutoML Regressor models.
7. Generate Predictions:
    • Generate predictions on new data using both models.
8. Visualize Results:
    • Visualize model performance, e.g., through plots or charts comparing MSE and R-squared scores.
```

V. SIMULATION RESULTS

RANDOM FOREST REGRESSOR SIMULATION RESULT



```
Best Parameters: {'regressor__max_depth': None, 'regressor__min_samples_split': 2, 'regressor__n_estimators': 200}
Cross-Validation R-squared Scores: [0.99768118 0.99982829 0.9998108  0.99909318 0.99676483]
Mean R-squared Score: 0.9986356560135892
Polynomial Features MSE: 770183.3459786888
Polynomial Features R-squared Score: 0.9996111178763422
Test Set MSE: 858579.9147669001
Test Set R-squared Score: 0.9995664845490001
```

The following hyperparameters were used to train the Random Forest model: max_depth=None, min_samples_split=2, and n_estimators=200. The grid search method was used to choose these hyperparameters in order to maximize the performance of the model.

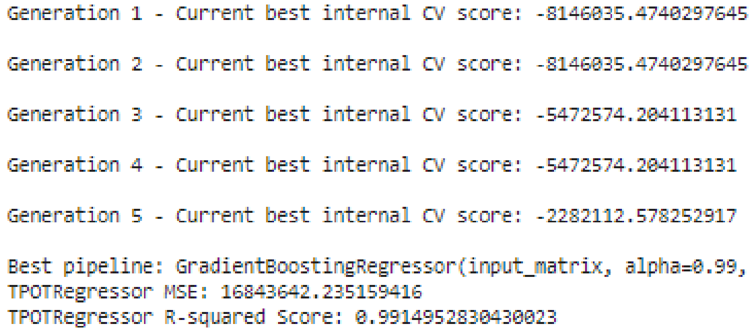
Strong predictive potential was demonstrated by the Random Forest model during cross-validation, as seen by its high R-squared scores over numerous folds. The target variable's variation may be explained by the model with an overall efficiency of around 0.999, as indicated by the mean R-squared score.

Polynomial characteristics were included in order to increase the predictive potential of the model. With a larger degree of variation described by the model, the mean squared error (MSE) decreased to 770,183.35 and the R-squared score climbed to 0.9996.

With an MSE of 858,579.91 and an R-squared score of 0.9996, the Random Forest model performed consistently when tested on the test set. These findings support the model's strong capacity to generalize to new data, indicating that it is a reliable predictor of pizza pricing.

All things considered, the Random Forest model performs well, identifying intricate correlations in the data and producing precise predictions.

TPOT SIMULATION RESULT



```
Generation 1 - Current best internal CV score: -8146035.4740297645

Generation 2 - Current best internal CV score: -8146035.4740297645

Generation 3 - Current best internal CV score: -5472574.204113131

Generation 4 - Current best internal CV score: -5472574.204113131

Generation 5 - Current best internal CV score: -2282112.578252917

Best pipeline: GradientBoostingRegressor(input_matrix, alpha=0.99,
TPOTRegressor MSE: 16843642.235159416
TPOTRegressor R-squared Score: 0.9914952830430023
```

Internal cross-validation (CV) score optimization was the goal of each iteration of the TPOT AutoML process as it advanced. Based on its capacity to reduce the mean squared error (MSE) during cross-validation, the top-performing pipeline was chosen. The optimal model was determined by the TPOT AutoML method to be a GradientBoostingRegressor pipeline after several rounds. Alpha, learning_rate, loss function, max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators, and subsample are among the hyperparameters that are included in this pipeline. The model's prediction performance was optimized by adjusting these hyperparameters. The TPOTRegressor obtained an R-squared score of around 0.9915 and an MSE of 16,843,642.24 when assessed on the test set. These findings show that, although having a somewhat higher MSE, the TPOT AutoML technique effectively trained a model that predicts pizza prices**.** On the other hand, the R-squared value indicates that the TPOTRegressor model has good predictive power, explaining around 99.15% of the variation in the target variable.

In conclusion, the TPOT AutoML technique produces a high-performing predictive model for pizza price prediction by efficiently automating the model selection and hyperparameter tweaking processes.

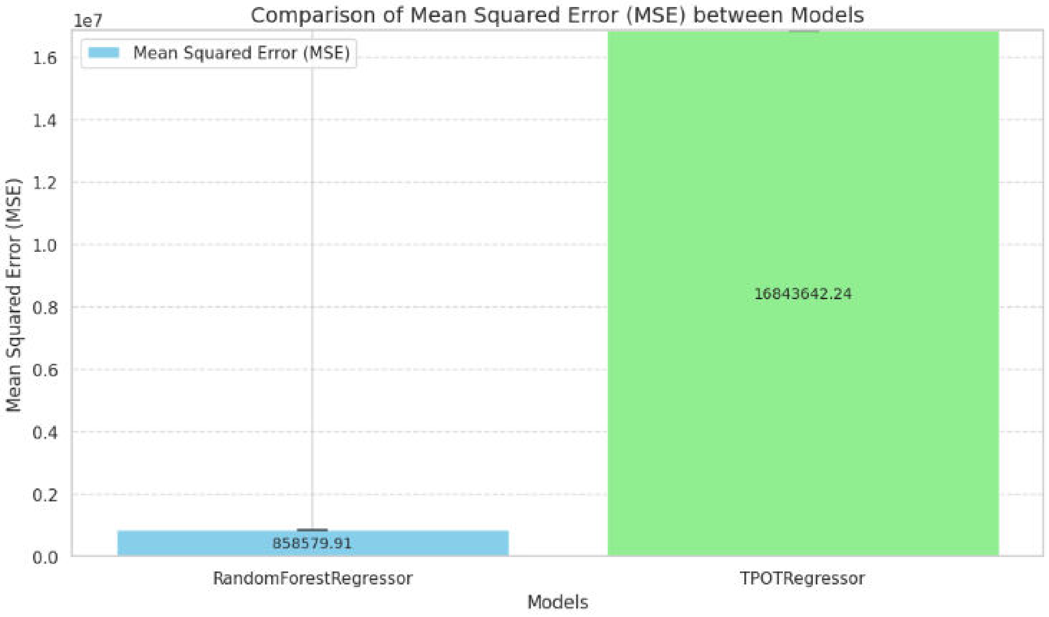


FIG1.1 COMPARISON OF MEAN SQUARED ERROR (MSE) BETWEEN RANDOM FOREST REGRESSOR AND TPOT AUTOML REGRESSOR

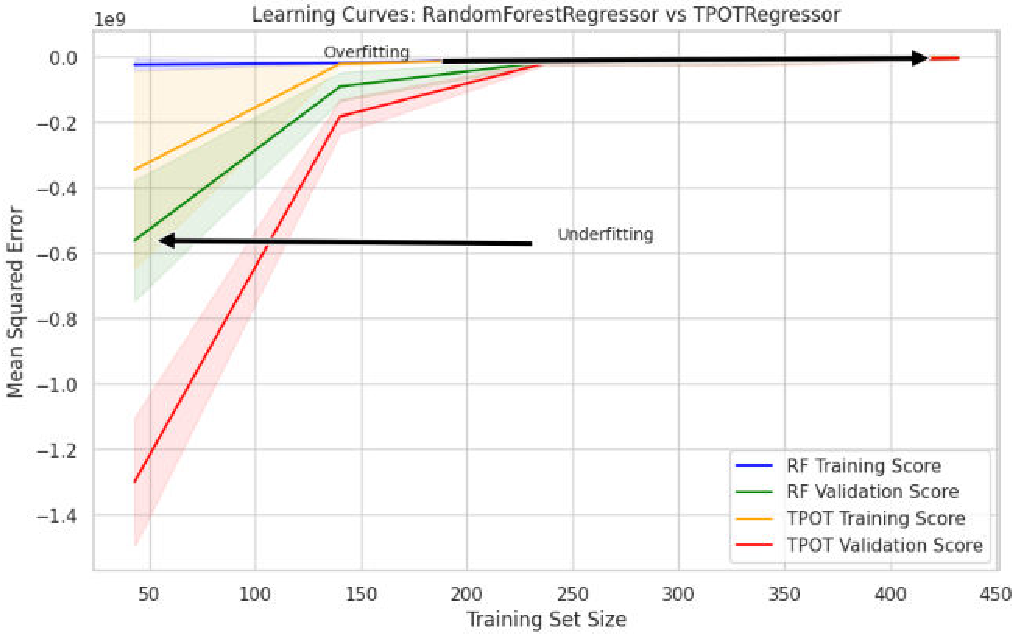


FIG1.2 COMPARISON OF LEARNING CURVES (TRAINING & VALIDATION SCORES) BETWEEN RANDOM FOREST REGRESSOR AND TPOT AUTOML REGRESSOR

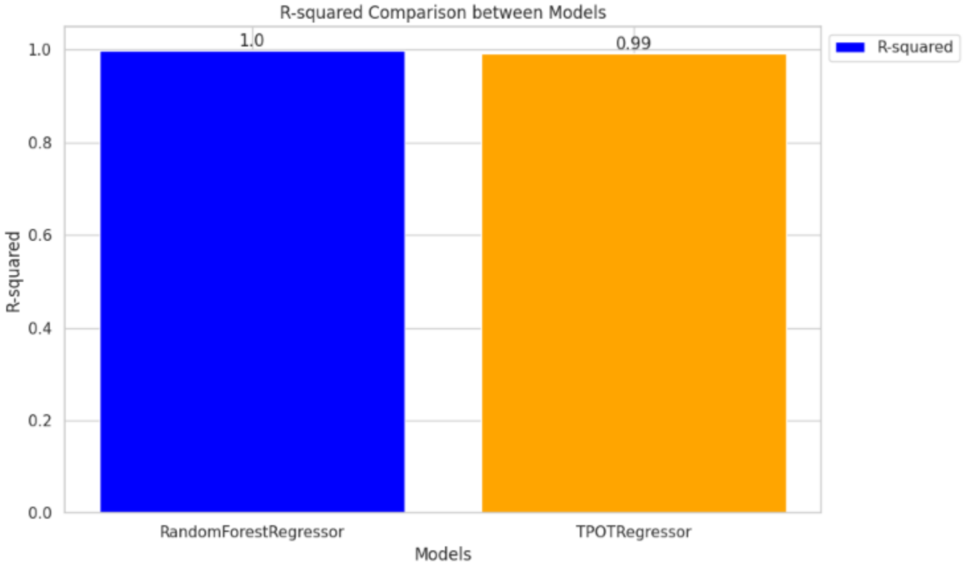


FIG1.3 COMPARISON OF R-SQUARED VALUES BETWEEN RANDOM FOREST REGRESSOR AND TPOT AUTOML REGRESSOR

VI. CONCLUSION AND FUTURE WORK

Several important conclusions and possible directions for further investigation are revealed by the comparison of the Random Forest and TPOT AutoML algorithms for pizza pricing prediction.

Conclusion:

The excellent R-squared values of the Random Forest and TPOT AutoML techniques indicate that they both performed well in forecasting pizza pricing. In comparison to the TPOT AutoML model, the Random Forest model produced a smaller Mean Squared Error (MSE) on the test set, suggesting somewhat higher prediction accuracy. TPOT AutoML demonstrated its capacity to automatically find and pick the optimal machine learning pipeline, utilizing cutting-edge methods like gradient boosting to produce outcomes that are competitive. The paper emphasizes how machine learning algorithms can effectively estimate intricate price structures, such those seen in the food business, and illustrates how crucial model selection and hyperparameter tweaking are to obtaining the best results.

Future Scope:

Examine Ensemble approaches: In order to integrate the advantages of various models and improve forecast accuracy, future research might examine the possible advantages of ensemble approaches like stacking or blending.
Include External Data: Including data from other sources, such customer preferences or economic indicators, might yield insightful information and increase the precision of forecasts for pizza prices.
Adjust Hyperparameters: Even more advanced models may result from experimenting with hyperparameter tuning methods, such as more thorough search approaches and cutting-edge optimization algorithms.
Handle Data Imbalance: Using strategies like oversampling or undersampling might reduce bias and enhance model generalization if the dataset shows signs of class imbalance or skewed distributions.

REFERENCES

[1] https://www.linkedin.com/pulse/pizza-market-detailed-analysis-current-scenario-etp0e/
[2] https://unacademy.com/content/railway-exam/study-material/static-gk/learn-all-about-the-pizza-market-in-india/

[3] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD InternationalConference on Knowledge Discovery and Data Mining, KDD ’16, San Francisco, CA, USA, 13–17 August 2016; Association forComputing Machinery: New York, NY, USA, 2016; pp. 785–794.
[4] Breiman, L. Random Forests. Mach. Learn. 2001,45, 5–32.
[5] https://www.sciencedirect.com/science/article/abs/pii/S0260877402002765
[6] https://www.researchgate.net/publication/311673623_Pizza_sales_forecasting_using_big_data_analysis
[7] https://www.mdpi.com/2304-8158/12/6/1347
[8] Darwish, A.; Ricci, M.; Zidane, F.; Vasquez, J.A.T.; Casu, M.R.; Lanteri, J.; Migliaccio, C.; Vipiana, F. Physical Contamination Detection in Food Industry Using Microwave and Machine Learning. Electronics 2022, 11, 3115.
[9] Fengou, L.C.; Liu, Y.; Roumani, D.; Tsakanikas, P.; Nychas, G.J.E. Spectroscopic Data for the Rapid Assessment of Microbiological Quality of Chicken Burgers. Foods 2022, 11, 2386.
[10] Özdoğan, G.; Lin, X.; Sun, D.W. Rapid and noninvasive sensory analyses of food products by hyperspectral imaging: Recent application developments. Trends Food Sci. Technol. 2021, 111, 151–165.
[11] https://www.kaggle.com/datasets/knightbearr/pizza-price-prediction

BIOGRAPHY

Prof. Abhishek Singh is working as an Associate Professor in the Computer Science and Engineering Department in Baderia Global Institute of Engineering and Management, Jabalpur, Madhya Pradesh. His area of interest includes Python, Machine Learning, Process Mining and Data Analytics.

Prof. Zohaib Hasan is working as an Associate Professor in the Computer Science and Engineering Department in Baderia Global Institute of Engineering and Management, Jabalpur, Madhya Pradesh. His area of interest includes CPP, Java, Python, Computer Networking and Machine Learning.

Prof. Nirdesh Jain is working as an Associate Professor in the Electronics and Communication Engineering Department in Baderia Global Institute of Engineering and Management, Jabalpur, Madhya Pradesh. His area of interest includes Signals and System, SQL and Data Analytics.