Report

1. Introduction:

The primary objective of this machine learning project was to develop accurate regression models for predicting a target variable. The dataset, sourced from Google Drive, contained various features that required careful preprocessing and analysis. The models selected for experimentation included Random Forest, Extra Trees, XGBoost, and CatBoost.

2. Data Preprocessing:

The initial step involved loading the dataset from Google Drive and scrutinizing its structure. The redundant 'datasetId' column was promptly identified and removed, as it did not contribute meaningfully to the predictive task. Further, the 'condition' feature, which contained categorical data, underwent one-hot encoding to transform it into a format suitable for training machine learning models. The dataset was then split into training and testing sets to facilitate model evaluation.

3. Feature Selection:

Recursive Feature Elimination (RFE) with Linear Regression was employed to identify the most significant features for predicting the target variable. The intention behind this feature selection process was to enhance model performance by focusing on the most relevant attributes. The selected features were subsequently utilized in the model training phase.

4. Model Training and Evaluation:

- Extra Trees:

  Strengths: Extra Trees models exhibit high predictive accuracy, making them suitable for complex datasets. They are also robust to outliers.

  Weaknesses: Extra Trees models may not perform as well with smaller datasets, as they rely on a larger number of decision trees for their strength.

- XGBoost:

  Strengths: XGBoost is renowned for its high performance and ability to handle missing data effectively. It is a popular choice in machine learning competitions.

Weaknesses: Sensitivity to outliers requires careful preprocessing, and optimal hyperparameter tuning is crucial for achieving peak performance.

- CatBoost:

  Strengths: CatBoost excels in handling categorical features, reducing the need for extensive preprocessing. It also performs well with minimal hyperparameter tuning.

  Weaknesses: The computational expense associated with training CatBoost models can be a limiting factor for large datasets.

5. Evaluation:

Ensemble of Random Forest, Extra Trees, and XGBoost models demonstrated improved predictive performance compared to individual models.

The Random Forest model displayed a Train Mean Squared Error (MSE) of 0.375, emphasizing its ability to fit the training data well.

Feature importance analysis provided insights into the critical variables influencing the target variable.

1. Model Comparison:

Random Forest and Extra Trees exhibited robust predictive capabilities, showcasing their effectiveness in capturing complex relationships within the data.

Feature importance analysis helped understand the contribution of individual features to the predictive task, providing valuable insights for model interpretation.

2. Ensemble Approach:

The ensemble approach involved combining predictions from Random Forest, Extra Trees, and XGBoost models using weighted averaging.

We performed a grid-search on all possible combinations of ensemble weights of these models and chose the best ones.

Strengths: The ensemble method leveraged the strengths of individual models, resulting in improved predictive performance.

Weaknesses: Fine-tuning the weights for the ensemble posed challenges, requiring iterative experimentation.

3. Hyperparameter Tuning (Grid Search for XGBoost):

Hyperparameter tuning was conducted for the XGBoost model using GridSearchCV to identify the optimal combination of hyperparameters.

Strengths: Hyperparameter tuning is essential for optimizing model performance and achieving better generalization.

Weaknesses: The computational cost associated with GridSearchCV can be high, especially for a large search space.

4. Results and Insights:

The ensemble approach outperformed individual models, underscoring the potential benefits of combining diverse algorithms for improved predictive accuracy.

Feature importance analysis provided insights into the variables crucial for predicting the target variable.

Challenges in fine-tuning ensemble weights highlighted the importance of a nuanced and iterative approach to model combination.

6. Conclusion:

The project successfully implemented and compared various regression models, demonstrating the effectiveness of ensemble techniques.

Feature selection and hyperparameter tuning played pivotal roles in achieving optimal model performance.

Insights gained during the experimentation phase provided valuable knowledge for future model development.

Team Name: suwi

Members:

Om Mittal (Team Leader) | +917003516965

Sounak Ghosh | +918910272026

Aryan Paul | +919748756167

Sushmit Dasgupta | +918100145170