PREDICTING COMPANY PROFIT USING REGRESSION ALGORITHMS

BY:- PRATYUSH INGALE
PUNE INSTITUTE OF COMPUTER TECHNOLOGY

INTRODUCTION

The purpose of this project is to develop a machine learning model that can predict the profit value of a company based on its R&D Spend, Administration Cost, and Marketing Spend. The dataset consists of information from 50 companies, including their respective profit values. The goal is to construct different regression algorithms, divide the data into training and testing sets, calculate various regression metrics, and choose the best model for accurate profit predictions.

DATA COLLECTION AND PREPROCESSING

The dataset containing R&D Spend, Administration Cost, Marketing Spend, and Profit values for 50 companies was obtained. Data preprocessing steps such as handling missing values, encoding categorical variables, and scaling features were performed to ensure data quality and model compatibility.

REGRESSION ALGORITHMS

In this project, we constructed different regression algorithms to predict company profit. The following regression algorithms were implemented:

Linear Regression

A simple and interpretable algorithm that assumes a linear relationship between the features and target variable.

Decision Tree Regression

A non-linear algorithm that constructs a tree-like model to make predictions.

Random Forest Regression

An ensemble algorithm that combines multiple decision trees to improve prediction accuracy.

TRAIN-TEST SPLIT

To evaluate the performance of the regression models, the dataset was divided into a training set and a testing set. The training set was used to train the models, and the testing set was used to assess their predictive capabilities. The data was split in a 80:20 ratio, with 80% used for training and 20% for testing.

REGRESSION METRICS

Different regression metrics were calculated to evaluate the performance of the models. The following metrics were used:

- Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values, providing a measure of the model's overall error.

- Mean Absolute Error (MAE): Computes the average absolute difference between predicted and actual values, giving an indication of the model's average prediction error.

- R-squared (R2) Score: Determines the proportion of the variance in the target variable that can be explained by the model. It indicates how well the model fits the data.

RESULTS AND MODEL SELECTION

After training and testing the regression models, the regression metrics were calculated. The metrics provided insights into the performance of each model. Based on the metrics, the best model was chosen.

- Linear Regression: MSE= 80926321.22295,
MAE = 6979.1522, R2 Score =0.90

- Decision Tree Regression: MSE = 400026479.25494, MAE = 13755.6639, R2 Score = 0.506

- Random Forest Regression: MSE = 72625008.623, MAE = 6437.497, R2 Score = 0.9103

RESULTS AND MODEL SELECTION

Among the implemented models, the Random Forest Regression demonstrated superior performance with the lowest 72625008.62306513. Thus, the Random Forest Regression is recommended as the most suitable algorithm for predicting the profit value of a company based on R&D Spend, Administration Cost, and Marketing Spend.

CONCLUSION

In this project, we successfully developed a machine learning model to predict the profit value of a company using different regression algorithms. By comparing various regression metrics, we determined the best model for accurate profit predictions. This model can be used to gain valuable insights and make informed decisions regarding company profitability based on R&D Spend, Administration Cost, and Marketing Spend. Further improvements and optimizations can be explored to enhance the model's performance and extend its applicability.

FUTURE GOALS

- Feature Engineering: Explore additional features or transformations of existing features that may have a stronger correlation with the target variable (profit). Feature engineering techniques such as polynomial features, interaction terms, or domain-specific feature engineering can be applied to improve the model's performance.

- Model Optimization: Experiment with different hyperparameter tuning techniques to optimize the performance of the selected regression model. Grid search, random search, or Bayesian optimization can be used to fine-tune the model and achieve better results.

- Model Ensemble: Investigate the possibility of creating an ensemble model by combining the predictions of multiple regression models. Techniques such as stacking, blending, or boosting can be employed to leverage the strengths of different models and improve prediction accuracy.

- Cross-Validation: Implement cross-validation techniques, such as k-fold cross-validation or stratified cross-validation, to obtain a more robust evaluation of the models. This helps to assess their generalization performance and ensure that the model's performance is consistent across different subsets of the data.

FUTURE GOALS

- Feature Importance Analysis: Conduct a thorough analysis of feature importance to identify the most influential variables in predicting profit. Techniques like permutation importance, feature importance from tree-based models, or SHAP (SHapley Additive exPlanations) values can provide valuable insights into which features have the most impact on the target variable.

- Model Deployment: Once a satisfactory model is developed, consider deploying it into a production environment. This may involve creating a user-friendly interface, implementing an API, or integrating the model into a larger system to automate profit predictions for new data.

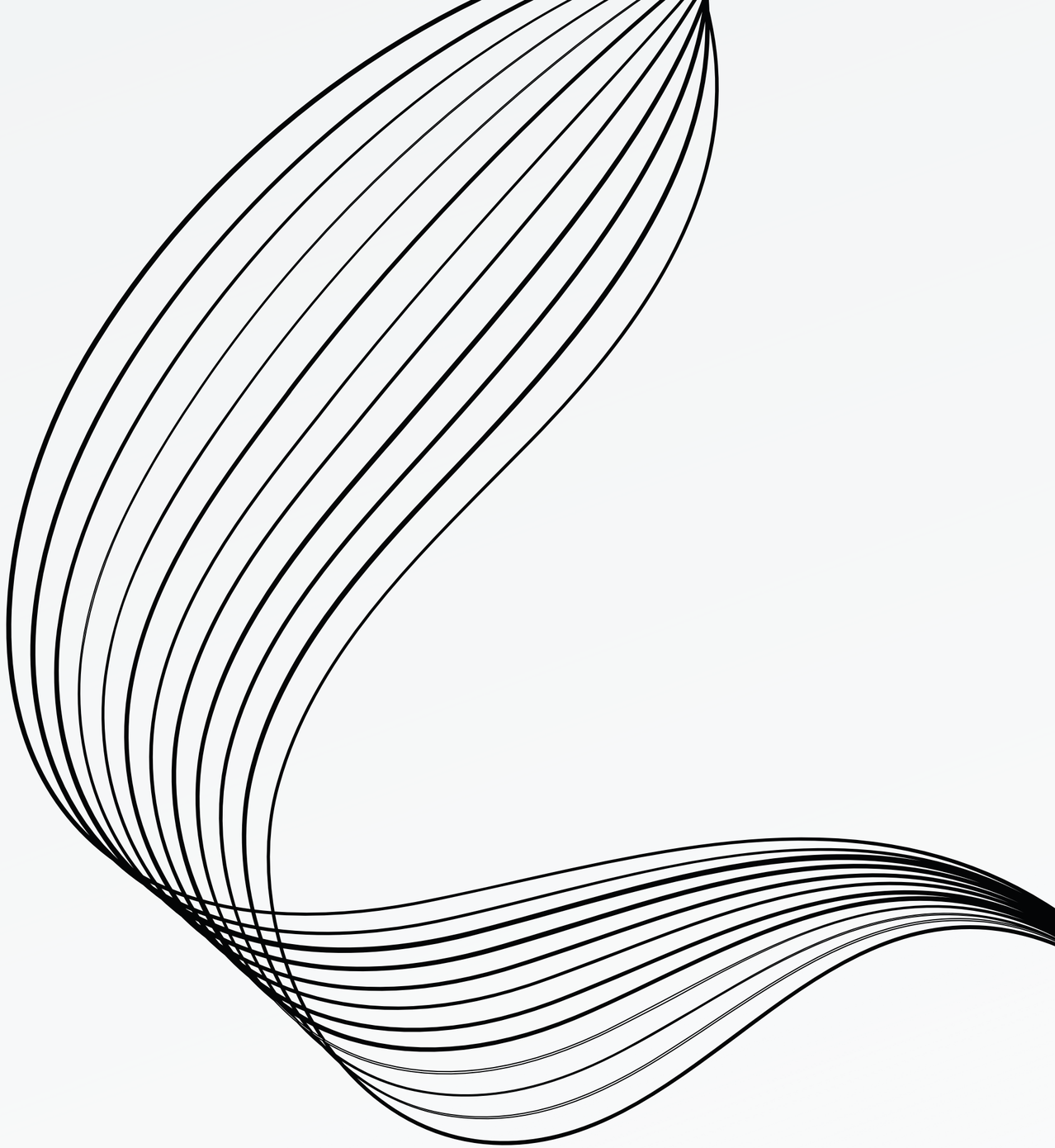
- Performance Monitoring and Updating: Continuously monitor the performance of the deployed model in the real-world setting. Collect feedback, track prediction errors, and retrain or update the model periodically to ensure its accuracy and relevance as new data becomes available.

- Business Insights: Analyze the model's predictions and the relationships between different features and profit to extract valuable business insights. Identify patterns, trends, or factors that contribute significantly to profit, which can guide strategic decision-making and resource allocation within the companies.

REFERENCES

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- Raschka, S., & Mirjalili, V. (2017). Python Machine Learning. Packt Publishing.
- Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.
- Sklearn Documentation: Scikit-learn: Machine Learning in Python. Retrieved from https://scikit-learn.org/stable/index.html
- Kaggle Datasets: Various datasets and resources for machine learning projects. Retrieved from https://www.kaggle.com/datasets

THANK YOU !



PRATYUSH INGALE