**Title: Predicting Company’s Profit using Machine Learning Regression Models**

**Kalavakuri Abhinay**

**Abstract:**

In today's competitive business landscape, accurately predicting a company's profit is crucial for informed decision-making and strategic planning. This abstract presents a machine learning regression model that utilizes a dataset consisting of R&D Spend, Administration Cost, Marketing Spend, and Profit columns to predict a company's profitability.

The proposed model employs a neural network that analyzes the relationship between the independent variables (R&D Spend, Administration Cost, and Marketing Spend) and the dependent variable (Profit). By analyzing the dataset, the model learns from historical data patterns and extracts valuable insights to make predictions on unseen or future data.

To develop the regression models, several steps are undertaken. Firstly, the dataset is preprocessed, including normalizing features, and addressing outliers. Various regression algorithms, such as Linear Regression, SGD Regression, Ridge Regression, Lasso Regression, Decision tree Regression ,Random Forest Regressor and Neural Networks are explored and compared to determine the best-fit model.

The model's performance is assessed through evaluation metrics such as root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE), and R-squared (R2) to gauge the accuracy and goodness of fit.

The results obtained from the regression model can provide valuable insights to company stakeholders and decision-makers. By understanding the impact of R&D Spend, Administration Cost, and Marketing Spend on Profit, companies can optimize their resource allocation, marketing strategies, and operational decisions to maximize profitability. Moreover, the model can help identify the key drivers that significantly influence the company's profit, enabling proactive measures to be taken to enhance financial performance.

In conclusion, this abstract proposes a neural network model that effectively predicts a company's profit using a dataset containing R&D Spend, Administration Cost, Marketing Spend, and Profit columns. The model holds the potential to provide valuable insights and support decision-making processes, ultimately assisting companies in achieving sustainable growth and profitability in a highly competitive business environment.

**Keywords:** Linear Regression, SGD Regression, Ridge Regression, Lasso Regression, Decision tree Regression, Random Forest Regressor, Neural Network, Root Mean Squared Error (RMSE), Mean squared error (MSE), mean absolute error (MAE), R-squared (R2).

**Table of Contents**

* Introduction
* Existing Method
* Proposed Method with Architecture
* Methodology
* Implementation
* Conclusion

**Introduction**

In today's data-driven world, understanding the relationship between variables and making accurate predictions is of utmost importance in various fields. Regression analysis, a fundamental statistical modeling technique, offers a powerful framework for analyzing and predicting continuous outcomes based on input variables. This report aims to explore the concept and application of regression models, providing a comprehensive understanding of their principles, assumptions, and interpretation.

Regression models enable us to examine the influence of one or more independent variables on a dependent variable, allowing us to quantify and analyze the nature of their relationship. By fitting a regression model to a dataset, we can uncover patterns, identify significant predictors, and estimate the impact of these predictors on the outcome of interest. Moreover, regression models provide a robust foundation for making predictions and understanding the variability inherent in the data.

**Existing Methods**

From the existing regression models like Linear Regression, SGD Regressor, Ridge Regression, Random Forest Regressor, Decision Tree Regressor, Lasso Regressor the metrics of the models on the testing set are tabulated below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | RMSE | MAE | MSE | R2 score |
| Linear Regressor | 0.24414928017169363 | 0.2086541702998223 | 0.05960887100835615 | 0.9340489707893037 |
| SGD Regressor | 0.3500958631440479 | 0.25663915605634374 | 0.12256711339057591 | 0.8643922097709342 |
| Ridge Regressor | 0.2514437460744679 | 0.2123801521476259 | 0.06322395743996148 | 0.9300492528477146 |
| Random Forest Regressor | 0.361722923610428 | 0.28484561408554976 | 0.13084347346527556 | 0.8552352763177198 |
| Decision Tree Regressor | 0.35699137174630186 | 0.26694303275458153 | 0.12744283950130628 | 0.8589977248610111 |
| Lasso Regressor | 1.0124401658635624 | 0.8269035942223661 | 1.0250350894538378 | -0.13409494229612418 |

From the table above, Linear Regressor model has the least error. Therefore, Linear Regressor is the preferred model for this type of dataset.

**Proposed Method with Architecture**

By using Neural Networks and the following architecture is implemented.

A screen shot of a computer code

Description automatically generated

A screen shot of a computer code

Description automatically generated

**Methodology**

1. **Data and Pre-processing**: The dataset consists of 50 rows and 4 columns (R&D Spend, Administration, Marketing Spend are features columns and Profit is the target column). The dataset contains Zero values for which the profits are non-zero, so removing them for appropriate predictions. The outliers are also removed and the data is then Normalized.
2. **Feature selection**: All the columns were used in predicting the profit of the company as per the question.
3. **Model Selection**: Various models like Linear Regression, SGD Regressor, Ridge Regression, Random Forest Regressor, Decision Tree Regressor, Lasso Regressor and Neural Networks are used and are selected based on the error metrics.
4. **Model Training**: The dataset is divided into 70% for training and 30% for testing which consists of 32 rows in training and 14 rows for testing sets respectively. Hyperparameters like activation=”relu” , optimizer=”adam”, loss=”mse”, epochs=600 are used in training the neural network. Regularization technique like early stopping of patience=6 is also used in monitoring the validation loss in neural network.
5. **Model Evaluation**: The metrics of neural network are evaluated on the testing set and are as follows MSE: 0.03709562599155429 , MAE: 0.16251022641238796, RMSE: 0.1926022481477158.
6. **Results and Discussion**: The error in Neural Network is significantly lower than the linear regression model and other models.

**Implementation Details**

**Language** : Python 3.9.16

**Libraries used (with versions):** NumPy (1.24.3), Pandas (2.0.1) ,Seaborn (0.12.2), Matplotlib (3.7.1) , sklearn-contrib-lightning (0.6.0) , Keras (2.10.0), jupyter (1.0.0), notebook (6.5.4)

**Conclusion**:

Neural networks have the advantage of being able to capture complex nonlinear relationships in the data, making them well-suited for regression tasks where the input features have intricate interactions. By adjusting the architecture, the number of hidden layers, and the number of neurons in each layer, neural networks can learn to approximate the underlying function mapping the input to the continuous output, thus performing regression.