Predicting Startup Profit Using Machine Learning

(1 MONTH - DATA SCIENCE INTERNSHIP)

SUBMITTED TO:- EXPOSYS DATA LABS

SUBMITTED BY- SHIVSHANKAR KUMAR

Table Of Contents:

ABSTRACT

This report explores the application of machine learning techniques to predict startup profitability. Various regression models are evaluated to determine the most accurate method for predicting profits based on key financial indicators. The study focuses on data preprocessing, exploratory data analysis, model training, and performance comparison. Additionally, feature selection techniques are utilized to enhance model efficiency. The insights gained from this analysis can assist entrepreneurs and investors in making data-driven business decisions.

INTRODUCTION

Startups operate in a highly competitive environment, where financial forecasting is crucial for growth and sustainability. Machine learning provides a data-driven approach to predict future profitability based on various business attributes. This report aims to analyse different regression models and determine the most suitable one for profit prediction. By leveraging historical data and advanced analytical techniques, businesses can make informed investment decisions. The integration of automated predictive systems enhances accuracy and reduces reliance on traditional forecasting methods.

EXISTING METHOD

Traditional methods of financial forecasting rely on manual analysis, rule-based approaches, or simple statistical models. These techniques often fail to capture complex nonlinear relationships between features and outcomes. Common approaches include:

- Basic financial trend analysis
- Linear regression with limited features
- Manual economic indicators evaluation
-

PROPOSED APPROACH

To overcome the limitations of existing methods, a machine learning-based predictive model was developed. The proposed architecture includes:

1. **Data Preprocessing**: Handling missing values, encoding categorical variables, and normalizing data.
2. **Exploratory Data Analysis (EDA)**: Pair plots, correlation heatmaps, and statistical summaries to understand feature relationships.
3. **Model Training & Evaluation**: Training multiple regression models and comparing performance based on Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score.
4. **Best Model Selection**: Choosing the model with the highest accuracy.

5. **User Prediction System**: A feature that allows users to input startup details and obtain a profit prediction.

METHODOLOGY

Data Collection

- The dataset "50_Startups.csv" was used, containing R&D Spend, Administration, Marketing Spend, State, and Profit columns.

Data Preprocessing

- Missing values were checked and handled.
- Categorical data (State) was one-hot encoded.
- The dataset was split into training and testing sets (80% training, 20% testing).

Exploratory Data Analysis (EDA)

- Pair-plots and correlation heatmaps were generated to analyse relationships between features.

Model Selection

- The following models were trained and evaluated:
    1. **Linear Regression**
    2. **Ridge Regression**
    3. **Lasso Regression**
    4. **Random Forest Regressor**
- Metrics such as MAE, MSE, RMSE, and R^2 Score were used for evaluation.

IMPLEMENTATION

Loading Data:

```
data = pd.read_csv('/content/50_Startups.csv')
```

Data Preprocessing:

```
if 'State' in data.columns:
    data = pd.get_dummies(data, columns=['State'], drop_first=True)
X = data.drop('Profit', axis=1)
y = data['Profit']
```

Splitting Data:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                    random_state=42)
```

Training Models:

```
models = {
    "Linear Regression": LinearRegression(),
    "Ridge Regression": Ridge(),
    "Lasso Regression": Lasso(),
    "Random Forest Regressor": RandomForestRegressor(random_state=42)
}
```

Evaluating Models:

```
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    r2 = r2_score(y_test, y_pred)
    print(f"{name}: R² Score = {r2}")
```

Best Model Selection:

```
best_model = results_df.loc[results_df['R² Score'].idxmax()]
print(f"Best Model: {best_model['Model']}")
```

CONCLUSION

This study demonstrated the effectiveness of machine learning in predicting startup profitability. Among the tested models, **Random Forest Regressor** emerged as the best-performing model with the highest accuracy. The developed system enables startups to estimate profitability based on their investments in R&D, administration, and marketing. Future improvements could include adding more features, fine-tuning hyperparameters, and deploying the model as a web-based tool for broader accessibility.