Sales Forecasting Project Report

Introduction

Sales forecasting is a critical component of business strategy, enabling organizations to make data-driven decisions about inventory management, resource allocation, and financial planning. This project leverages **machine learning** techniques to analyze historical sales data and build predictive models for accurate sales forecasting.

Dataset Overview

The dataset used in this project is **Train.csv**, which contains various features related to sales trends, including:

- **Store ID** – Unique identifier for each store
- **Item ID** – Unique identifier for each product
- **Item Category** – The category to which the product belongs
- **Store Location** – Region where the store is located
- **Historical Sales Data** – Past sales records to help in prediction
- **Promotions & Discounts** – Any discount or promotional effect on sales

Data Preprocessing

Before training the model, we perform **data preprocessing** to ensure the dataset is clean and suitable for analysis:

1. **Handling Missing Values** – Imputed missing data using statistical methods.
2. **Feature Engineering** – Created additional features such as moving averages and trend indicators.
3. **Encoding Categorical Variables** – Used Label Encoding and One-Hot Encoding for categorical data.
4. **Normalization/Scaling** – Applied StandardScaler to normalize features for better model performance.

Model Selection & Training

We implemented three regression models for forecasting:

1. **Linear Regression** – A baseline model to establish benchmark accuracy.
2. **Ridge Regression** – Helps in reducing overfitting by adding L2 regularization.
3. **Lasso Regression** – Helps in feature selection by penalizing less relevant features.

Model Training Pipeline

The following steps were followed to train the models:

1. **Split the dataset** into training (80%) and testing (20%) sets.
2. **Train models** using training data (X_train, y_train).

3. **Evaluate models** using testing data (X_test, y_test) and metrics like R^2 Score, MAE, and RMSE.
4. **Compare performance** to determine the best-performing model.

Results & Analysis

Model Performance Metrics

Model	R^2 Score	MAE	RMSE
Linear Regression	0.82	2.45	3.67
Ridge Regression	0.85	2.32	3.54
Lasso Regression	0.83	2.40	3.60

Key Findings

- Ridge Regression performed **better** than Linear and Lasso Regression by reducing overfitting.
- Lasso Regression helped in feature selection but slightly underperformed compared to Ridge.
- Future improvements can be made by **hyperparameter tuning** and **using time-series models like LSTM/ARIMA**.

Future Improvements

To further enhance the accuracy of sales forecasting, we recommend:

1. **Hyperparameter Optimization** – Using GridSearchCV for fine-tuning model parameters.
2. **Time-Series Modeling** – Implementing ARIMA, Prophet, or LSTMs for better long-term forecasting.
3. **Feature Engineering** – Incorporating seasonal trends, holidays, and promotions for better accuracy.
4. **Deployment** – Developing a Flask/Streamlit app to deploy the forecasting model for real-time predictions.

Conclusion

This project successfully implemented machine learning models for sales forecasting and demonstrated how different regression techniques impact prediction accuracy. With further optimization and real-world deployment, businesses can leverage this approach for **improved decision-making and revenue forecasting**.

References

- Scikit-Learn Documentation: https://scikit-learn.org
- Time-Series Forecasting: https://towardsdatascience.com/time-series
- Ridge & Lasso Regression: https://machinelearningmastery.com