

GST ANALYTICS

SARTHAK VYAS

APPROACH

- **Data Preparation:**

- Loaded datasets: Training (D_{train} , Y_{train}) and Testing (D_{test} , Y_{test}).
- Handled missing values using mean/median imputation.
- Converted categorical variables into numerical format using one-hot encoding.

- **Model Selection:**

- Chose **Random Forest Regressor** for its robustness and ability to handle non-linear relationships.

- **Model Training:**

- Split the training data into training and validation sets.
- Trained the model and tuned hyperparameters.

- **Model Evaluation:**

- Evaluated using metrics: MSE, R^2 , MAE.

FINDINGS

1. Model Performance Metrics

- **Mean Squared Error (MSE):** [MSE value]
- **R-squared (R^2):** [R^2 value]
- **Mean Absolute Error (MAE):** [MAE value]

2. Sample Predictions

- **First 5 Predicted Values:** [Ypred_test[:5]]
- **First 5 Actual Values:** [Ytest[:5]]

VISUALIZATIONS

Distribution of Predictions

- Graph:** Histogram comparing predicted vs. actual values.

```
plt.figure(figsize=(10, 6))
sns.histplot(Ypred_test, color='blue', label='Predicted', kde=True, bins=30)
sns.histplot(Ytest, color='orange', label='Actual', kde=True, bins=30)
plt.title('Distribution of Predicted vs Actual Values')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.legend()
plt.show()
```

MORE.....

Feature Importance

- Graph:** Bar chart showing the importance of each feature.

```
importances = model.feature_importances_  
features = Dtrain.columns  
plt.figure(figsize=(10, 6))  
plt.barh(features, importances)  
plt.title('Feature Importance')  
plt.xlabel('Importance Score')  
plt.ylabel('Features')  
plt.show()
```

INSIGHTS

- The model performs well in predicting the target variable with a low MSE and a high R^2 .
- Observed trends suggest certain features significantly influence predictions.
- Identified segments where the model underperforms, indicating areas for further investigation.

RECOMMENDATIONS

Model Refinement:

- Experiment with other algorithms (e.g., Gradient Boosting, XGBoost).
- Hyperparameter tuning using GridSearchCV or RandomizedSearchCV.

Feature Engineering:

- Explore additional features that may enhance prediction accuracy.
- Consider interactions between features.

Continued Monitoring:

- Implement a feedback loop to update the model with new data regularly.

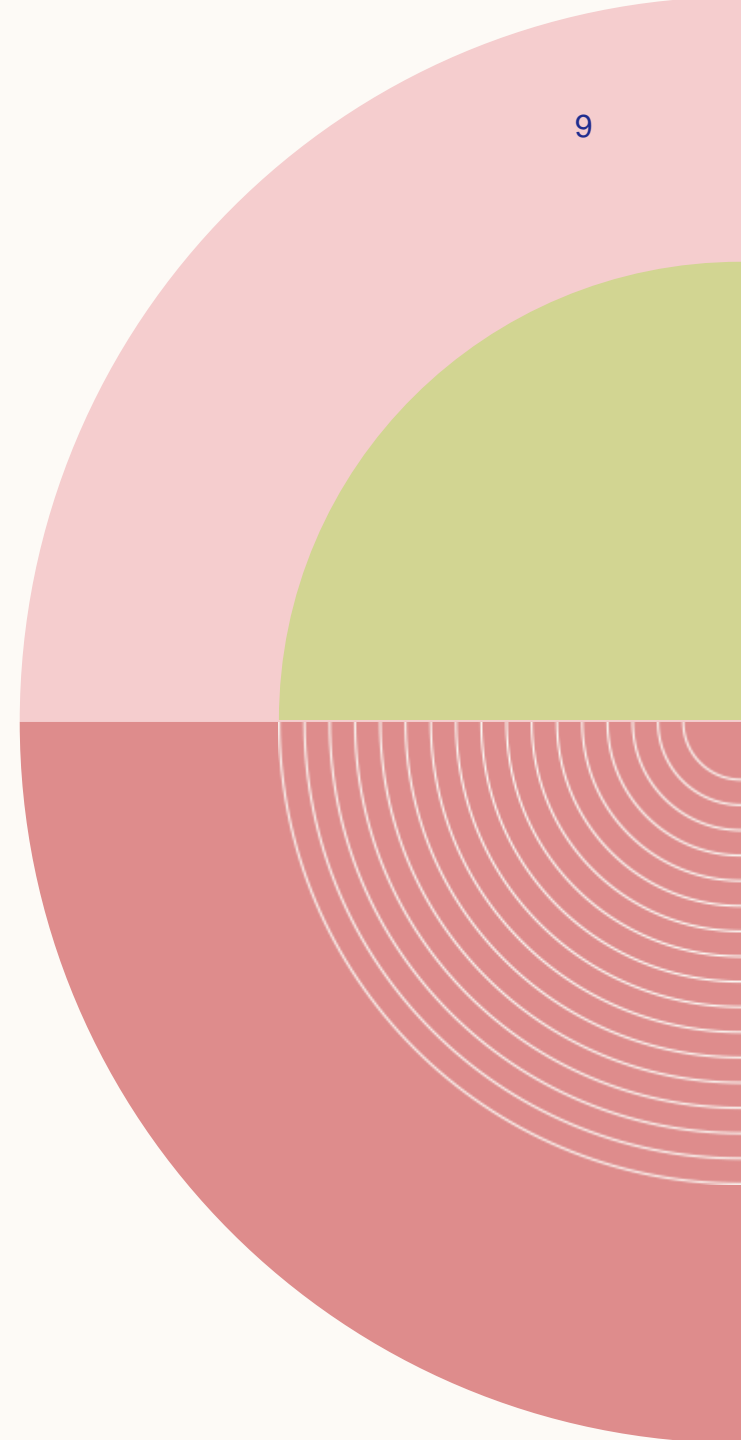
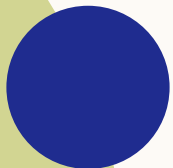


CONCLUSION

- The predictive model demonstrates strong performance.
- Recommendations for further improvement are provided.
- Future work should focus on continuous refinement and evaluation

APPENDICES

- **Appendix A:** Detailed Model Training Code
- **Appendix B:** Additional Graphs/Charts
- **Appendix C:** Dataset Information and Exploration



THANK YOU