# Machine Learning Assignment Final Report

*Kalhara J.A.K. - 214097U*

GitHub - https://github.com/Kalhara-JA/ML-Assignment-214097U

Demo App - https://lk-tourism-arrival-predictor.streamlit.app/

## Title

Forecasting Monthly Tourist Arrivals to Sri Lanka Using Traditional Machine Learning

## 1. Problem Definition and Dataset Collection

### 1.1 Problem Definition

Tourism is a critical sector for Sri Lanka. Accurate short-term forecasts of monthly tourist arrivals are useful for staffing, planning, and policy-level decision support in tourism-related services.

This project addresses the following problem:

- Predict Sri Lanka's next monthly tourist arrival counts using historical monthly arrivals data.

This is a **supervised regression** task.

### 1.2 Dataset Source and Collection

- Dataset name: Sri Lanka Monthly Tourist Arrivals (2016-2025)
- Source owner: Sri Lanka Tourism Development Authority (SLTDA)
- Public source: https://www.sltda.gov.lk/en/statistics

Source pages used:

- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2025
- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2024
- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2023
- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2022
- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2021
- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2020
- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2019
- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2018
- https://www.sltda.gov.lk/en/monthly-tourist-arrivals-reports-2017

Local dataset artifact:

- `data/raw/sri_lanka_tourism_monthly_arrivals_2016_2025.csv`

### 1.3 Features and Target Variable

Raw fields:

- `year`
- `month`
- `month_name`
- `date`
- `arrivals` (target variable)
- `source_url`

Dataset summary:

- Number of rows: 120 monthly observations
- Date range: 2016-01-01 to 2025-12-01
- Missing values: 0

### 1.4 Preprocessing

Preprocessing and feature engineering steps:

- Sort by date.
- Create seasonal encodings:
  - `month_sin`
  - `month_cos`
- Create lag features:
  - `lag_1`, `lag_2`, `lag_3`, `lag_12`
- Create rolling features:
  - `rolling_3`, `rolling_6`
- Create trend features:
  - `pct_change_1`, `pct_change_12`
- Replace infinite values from percentage changes with nulls.
- Drop rows with null feature values after lag/rolling creation.

After feature engineering, usable rows were 103.

### 1.5 Ethical Data Use

- Data is publicly available aggregate statistics.
- No personal or sensitive individual-level data is used.
- Forecasts are used for academic and planning support context, not automated policy decisions.

## 2. Selection of Traditional Machine Learning Algorithms

Only traditional ML models were used:

- Ridge Regression

- SVR (RBF)
- Random Forest Regressor

Why these models:

- Ridge provides a robust linear baseline.
- SVR captures non-linear relationships with kernel methods.
- Random Forest captures non-linear and interaction effects with ensemble trees.

This model set provides a meaningful baseline comparison while satisfying the "traditional ML only" rule.

# 3. Model Training and Evaluation

## 3.1 Train/Validation/Test Split

Chronological split (time-series safe):

- Train: 79 rows
- Validation: 12 rows
- Test: 12 rows

No random shuffling was used, preventing temporal leakage.

## 3.2 Hyperparameter Tuning

Each model used GridSearchCV with:

- CV strategy: `TimeSeriesSplit(n_splits=5)`
- Selection score: negative RMSE in CV
- Final model choice: lowest validation RMSE

Best model selected:

- Random Forest with parameters:
  - `n_estimators=200`
  - `max_depth=None`
  - `min_samples_split=5`
  - `min_samples_leaf=1`

## 3.3 Metrics

Metrics used:

- MAE
- RMSE
- MAPE
- R2

## 3.4 Results

From `outputs/model_comparison.csv`:

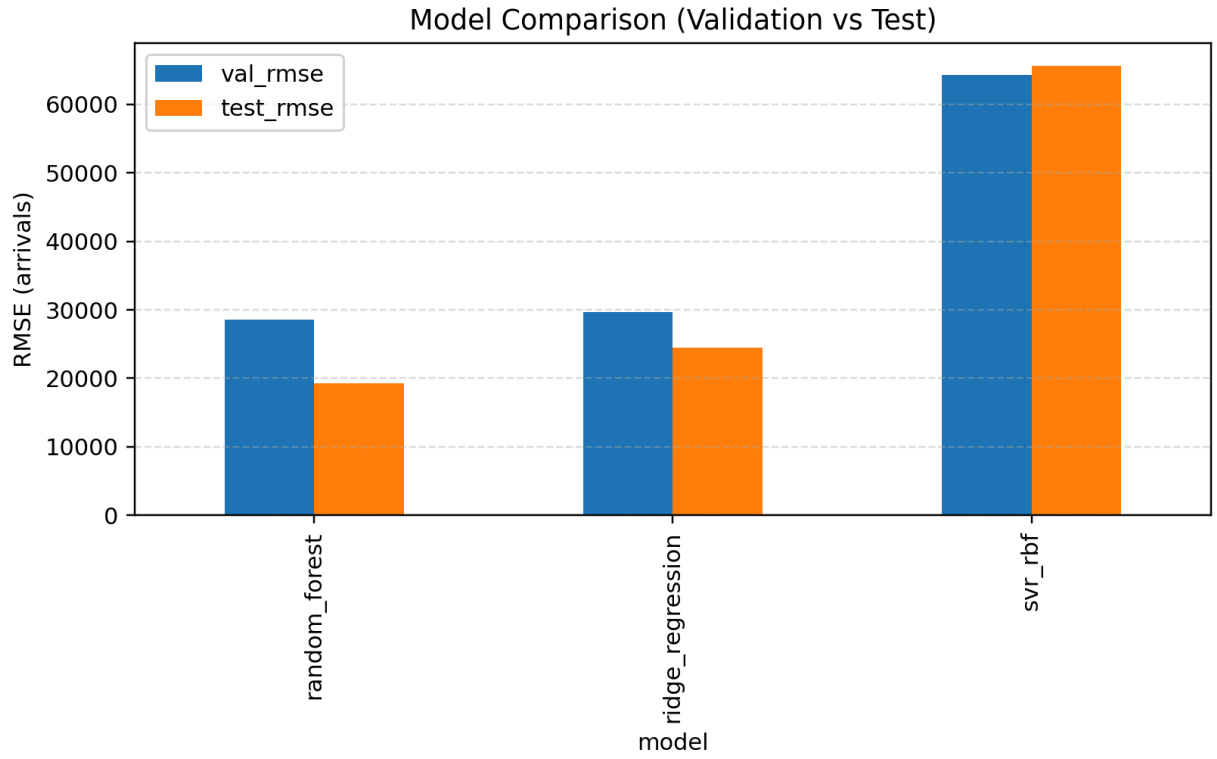
- Random Forest
    - Validation RMSE: 28,589.68
    - Test MAE: 15,864.72
    - Test RMSE: 19,255.20
    - Test MAPE: 10.55%
    - Test R2: 0.842
- Ridge Regression
    - Validation RMSE: 29,663.09
    - Test RMSE: 24,506.43
    - Test R2: 0.744
- SVR (RBF)
    - Validation RMSE: 64,296.32
    - Test RMSE: 65,627.57
    - Test R2: -0.837

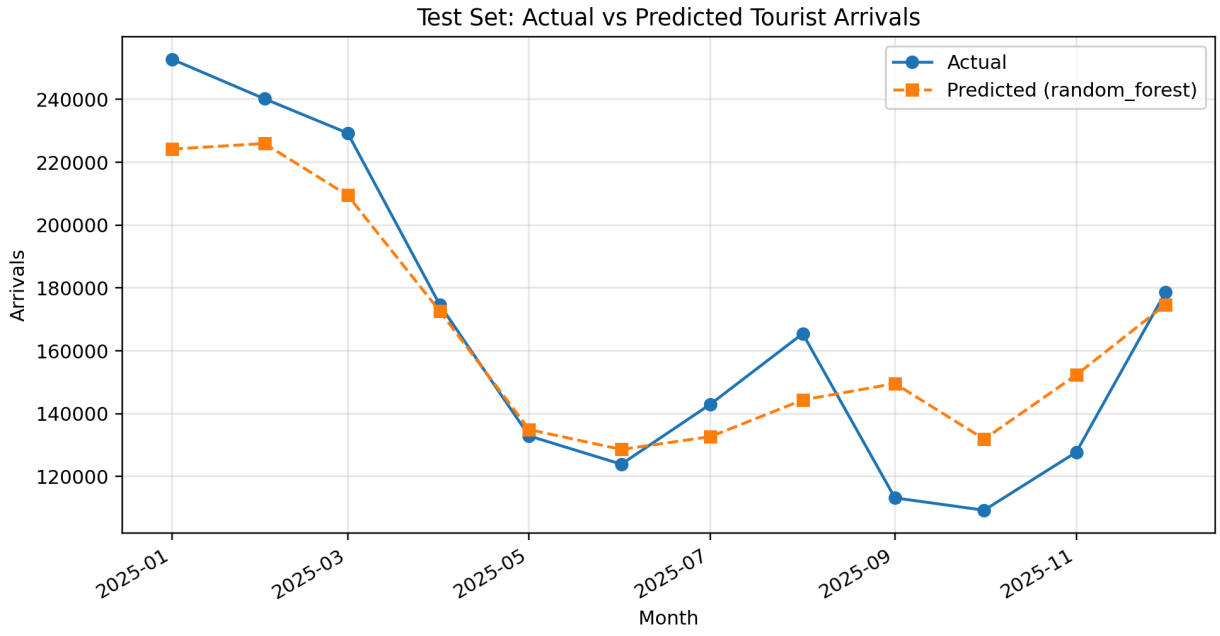
Interpretation:

- Random Forest clearly outperformed other models on test RMSE and R2.
- SVR underperformed likely due to data scale/structure and regime changes.

## 3.5 Evaluation Artifacts

- `outputs/model_comparison.csv`
- `outputs/best_model_metrics.json`
- `outputs/test_predictions.csv`
- `outputs/figures/model_comparison_rmse.png`
- `outputs/figures/test_actual_vs_predicted.png`




# 4. Explainability and Interpretation

Explainability methods used:

- SHAP (SHapley Additive exPlanations)

Top influential features (SHAP):

1. `lag_1`
2. `lag_12`
3. `rolling_3`
4. `pct_change_12`
5. `month`

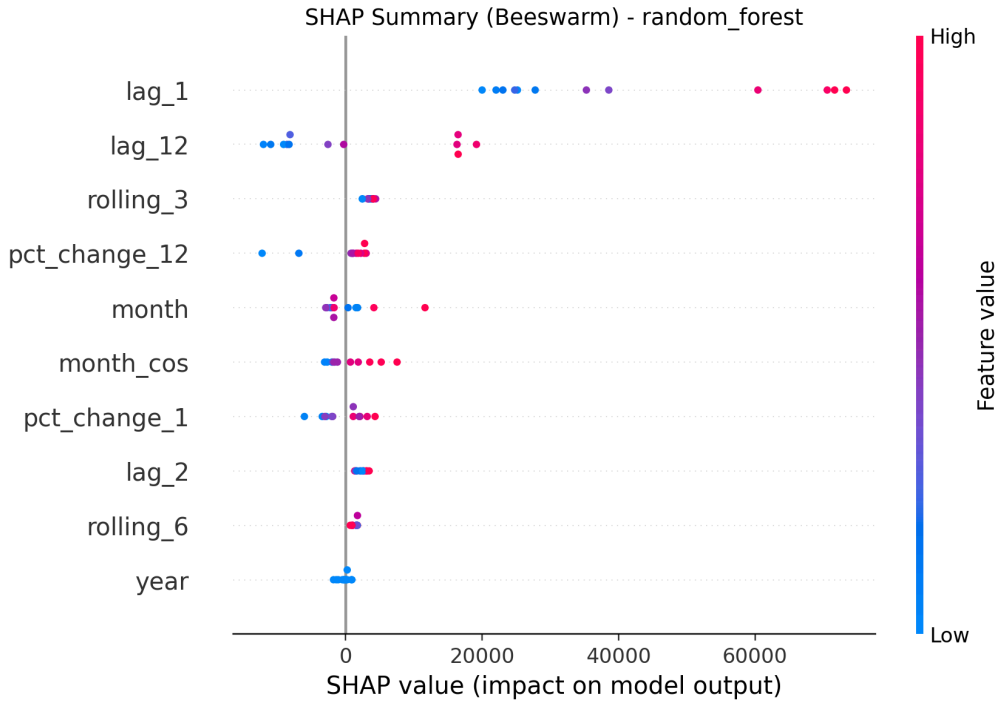
Interpretation:

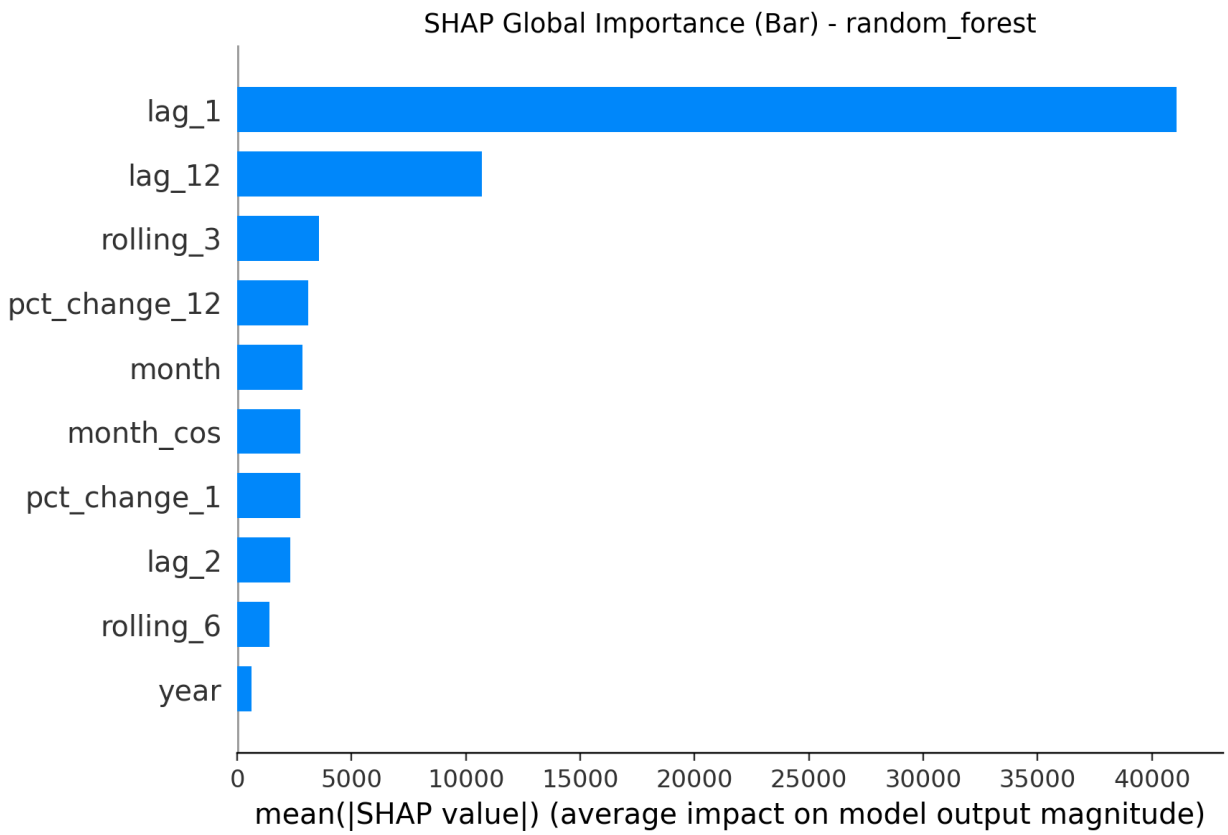
- `lag_1` indicates strong short-term temporal dependency.
- `lag_12` captures annual seasonality.
- Change-based features capture trend shifts.

This is consistent with domain behavior in monthly tourism demand data.

Explainability artifacts:

- `outputs/explainability/shap_feature_importance.csv`
- `outputs/figures/shap_summary_beeswarm.png`
- `outputs/figures/shap_summary_bar.png`






# 5. Critical Discussion

## 5.1 Limitations

- Limited sample size for time-series ML.
- Structural break during COVID period (2020-2021).
- No external drivers included (e.g., exchange rates, air capacity, geopolitical events).

## 5.2 Data Quality and Bias Risks

- Official aggregated data may still contain revisions/delays.
- Pandemic years can bias models trained on mixed regimes.
- National-level totals hide heterogeneity across source markets.

## 5.3 Real-World and Ethical Considerations

- Forecasts can support planning but should not be treated as guaranteed values.
- Decision-makers should combine model output with current policy and market intelligence.
- Regular retraining is required as new data becomes available.

# 6. Report Quality and Technical Clarity

Project is structured with reproducible scripts and documented outputs:

- Dataset preparation: `src/step1_prepare_dataset.py`
- Training/evaluation: `src/train_and_evaluate.py`
- Explainability: `src/explain_model.py`
- front-end: `app.py`

Run sequence:

1. `python src/step1_prepare_dataset.py`
2. `python src/train_and_evaluate.py`
3. `python src/explain_model.py`
4. `streamlit run app.py`

# 7. Front-End Integration

A Streamlit app is implemented for user-facing model interaction.

Implemented features:

- Forecast horizon selection (1-12 months)
- Scenario override for latest observed value
- Loading/progress feedback while generating forecasts
- Forecast table and chart
- CSV export of forecast results
- SHAP explainability tab
- About section with model metric definitions and usage notes

Files:

- App: `app.py`
- notes: `docs/04_frontend.md`

# Conclusion

This project successfully applies traditional machine learning to a Sri Lankan public dataset, compares multiple baseline models, explains model behavior with XAI methods, and deploys a front-end for interactive forecasting. The best-performing model is Random Forest, with strong test performance (RMSE 19,255; R2 0.842), while acknowledging limitations from data size and structural shocks.

# Appendix: Key Output Files

- `data/raw/sri_lanka_tourism_monthly_arrivals_2016_2025.csv`
- `outputs/step1_dataset_summary.json`
- `outputs/model_comparison.csv`

- outputs/best_model_metrics.json
- outputs/test_predictions.csv
- outputs/models/best_model.joblib
- outputs/explainability/shap_feature_importance.csv
- outputs/figures/model_comparison_rmse.png
- outputs/figures/test_actual_vs_predicted.png
- outputs/figures/shap_summary_beeswarm.png
- outputs/figures/shap_summary_bar.png