BIG DATA ANALYSIS AND PROJECT

Assignment 1 – Part c

Milestone 1C

MODELING AND RESULTS

Kishaiyan Vellaichamy Thangaraj
a1819309

INTRODUCTION:

Following extensive research, a review of numerous academic papers, and discussions with my peers, I have redirected the focus of my analysis. Initially, my goal was to predict the percentage of government expenditure allocated to mental health across various countries without specific geographical boundaries. However, due to the variability in data quality and types across different countries, generalizations cannot be accurately made. Consequently, my analysis now concentrates on predicting the Australian government's expenditure on mental health. This expenditure can be influenced by various factors such as drug use, unemployment, and social media. The analysis utilizes publicly available data from sources including the, WHO, Stat counter Global, the Australian Bureau of Statistics, and Statista.

Prediction Model:

The objective of this analysis is to predict changes in mental health expenditure for the Australian government, influenced by factors such as drug use, unemployment, and social media. This constitutes a regression problem, as it involves predicting a continuous variable—namely, the government's expenditure on mental health—based on multiple independent variables. The target variable in this study is the mental health expenditure of the Australian government.

Data Pre-Processing:

Although the datasets originate from reliable sources and thus exhibit minimal discrepancies, there were few NaN or null values, which were subsequently imputed to maintain data integrity prior to model training. This step is essential, as every data point is valuable for the analysis. Furthermore, data preprocessing included normalization to ensure consistency across different scales and handling of any missing values to prevent potential biases in the model. Feature selection was employed to identify the most relevant features influencing the target variable, thereby reducing data noise through techniques such as correlation heatmaps. This process not only decreased the dimensionality of the dataset but also enhanced the model's performance and interpretability by concentrating on the key factors driving mental health expenditure.

Model Selection:

The models selected for this analysis include various regression-based techniques: Linear Regression, Ridge Regression, Lasso Regression, ElasticNet, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Support Vector Regression (SVR). These models were chosen to leverage their diverse strengths in handling different aspects of the data. Linear Regression provides a straightforward approach, while Ridge and Lasso Regression offer regularization to prevent overfitting. ElasticNet combines both Ridge and Lasso penalties for enhanced flexibility. Decision Trees and Random Forests allow for complex, non-linear relationships, while Gradient Boosting enhances predictive performance through iterative refinement. Support Vector Regression (SVR) contributes by modeling non-linear relationships with high-dimensional data. Together, these models facilitate a comprehensive evaluation of factors influencing mental health expenditure and ensure robust predictions.

Methodology:

The methodology for this analysis involved several key steps to ensure robust model performance and accurate predictions. Initially, data was collected from various reliable sources and integrated to create a comprehensive dataset with relevant features and target variables. A preprocessing pipeline was established to manage data cleaning, normalization, and imputation, which helped maintain consistency and prevent data leakage by integrating these steps within the model training process. The dataset was then divided into training and testing sets using a 70-30 split ratio to facilitate effective model training and evaluation. To enhance model performance and minimize noise, irrelevant columns were dropped, and the forward feature selection method was used to identify the most significant features affecting the target variable. A range of regression-based models, including Linear Regression, Ridge Regression, Lasso Regression, ElasticNet, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Support Vector Regression (SVR), were evaluated. Hyperparameters for each model were optimized through cross-validation, and the models were assessed using the R^2 metric to measure the proportion of variance explained. The best-performing models were selected based on their R^2 scores, ensuring the predictions were both accurate and reliable.

```
--- Linear Regression ---
Best Params: {}
R² Score: 0.8372237126325058

--- Ridge ---
Best Params: {'regressor__alpha': 1.0}
R² Score: 0.8652259339273544

--- Lasso ---
Best Params: {'regressor__alpha': 0.1}
R² Score: 0.9772333337206606

--- ElasticNet ---
Best Params: {'regressor__alpha': 1.0, 'regressor__l1_ratio': 0.8}
R² Score: 0.858620298184438

--- Decision Tree ---
Best Params: {'regressor__max_depth': 5}
R² Score: 0.6737717345397689

--- Random Forest ---
Best Params: {'regressor__max_depth': 5, 'regressor__n_estimators': 50}
R² Score: 0.7520362071855063

--- Gradient Boosting ---
Best Params: {'regressor__learning_rate': 0.1, 'regressor__n_estimators': 100}
R² Score: 0.7611113939720069

--- Support Vector Machine ---
Best Params: {'regressor__C': 10.0, 'regressor__gamma': 'auto'}
R² Score: -0.1242103955197067
```

Figure 1 GridSearchCV Results for various models with r2 score

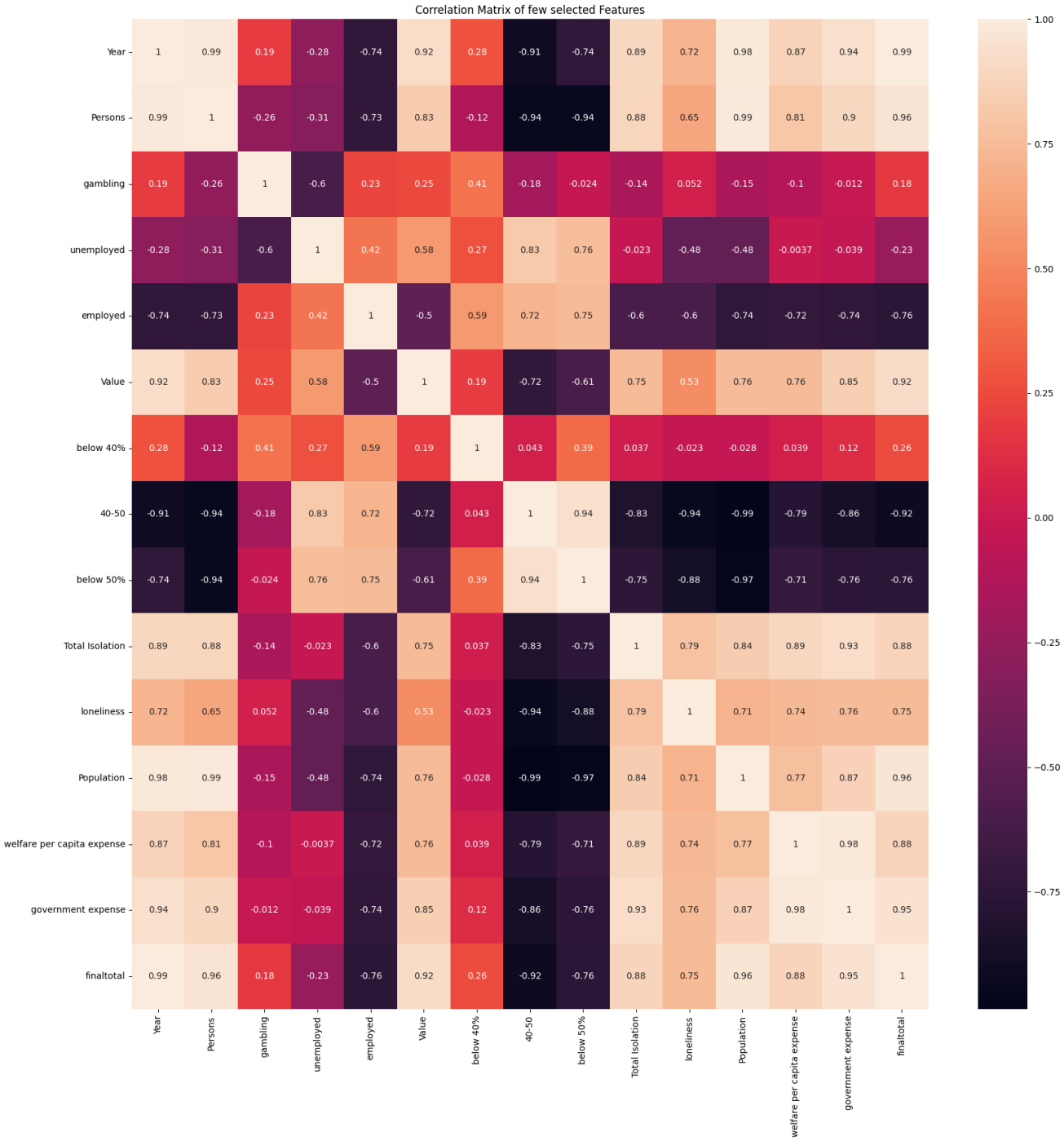

Figure 2 Heat Map for selected features from the dataset

Result:

The results indicate that Linear Regression with Lasso regularization emerged as the best-performing model, achieving an R^2 score of 0.97. This high R^2 score signifies that the model explains 97% of the variance in mental health expenditure, demonstrating a strong fit and high predictive accuracy. The Lasso regularization not only helped in improving the model's performance by addressing potential overfitting but also contributed to feature selection by penalizing less significant predictors. This resulted in a more streamlined and interpretable model that focuses on the most impactful factors influencing mental health expenditure. The exceptional R^2 score underscores the model's effectiveness in capturing the underlying trends and relationships within the data, making it a valuable tool for predicting future changes in mental health expenditure.

Overall, this model provides a robust framework for understanding and forecasting expenditure trends based on key influencing factors such as drug use, unemployment, and social media impact. Its high accuracy and interpretability make it especially useful for policymakers and stakeholders who need to make informed decisions about resource allocation and strategic planning. By leveraging this model, decision-makers can gain insights into how changes in the influencing factors might affect future expenditure, allowing for more proactive and data-driven approaches to mental health funding and intervention strategies.

References:

- *Expenditure - Mental Health* (no date) *Australian Institute of Health and Welfare*. Available at: https://www.aihw.gov.au/mental-health/topic-areas/expenditure (Accessed: 21 July 2024).
- *Social Determinants of Health* (no date) *Australian Institute of Health and Welfare*. Available at: https://www.aihw.gov.au/reports/australias-health/social-determinants-of-health (Accessed: 21 July 2024).
- Hughes, C. (2023) *Social media use in Australia 2022*, *Statista*. Available at: https://www.statista.com/statistics/680201/australia-social-media-penetration/ (Accessed: 21 July 2024).
- *Prevalence and impact of mental illness - mental health* (no date) *Australian Institute of Health and Welfare*. Available at: https://www.aihw.gov.au/mental-health/overview/prevalence-and-impact-of-mental-illness#moreinfo (Accessed: 21 July 2024).
- Parslow, R.A. and Jorm, A.F. (2000) 'Who uses Mental Health Services in Australia? an analysis of data from the National Survey of Mental Health and Wellbeing', *Australian & New Zealand Journal of Psychiatry*, 34(6), pp. 997–1008. doi:10.1080/000486700276.
- Andrews, G. (1999) *The mental health of Australians*. Canberra: Mental Health Branch, Commonwealth Department of Health and Aged Care.
- Tate, A.E. *et al.* (2020) 'Predicting mental health problems in adolescence using machine learning techniques', *PLOS ONE*, 15(4). doi:10.1371/journal.pone.0230389.
- Pandey, M. *et al.* (2021) 'Mental health prediction for juvenile using Machine Learning Techniques', *SSRN Electronic Journal* [Preprint]. doi:10.2139/ssrn.3867291.