BIG DATA ANALYSIS AND PROJECT

Assignment – Part D

Milestone 1D

REPORT

Kishaiyan Vellaichamy Thangaraj
a1819309

INTRODUCTION:

The research question evolved from a broad analysis of a country's increasing expenditure in the health sector, particularly focusing on mental health, to a more refined investigation centered on Australia's specific expenditures on mental health. The analysis considered various factors impacting individual mental health, including the government's historical data on GDP, health expenditure, and mental health expenditure. Additional data, such as social media penetration, COVID-19 statistics, and other relevant factors, were sourced from Statista, WHO, and Stat Counter Global. The answer for the question was the mental health expenditure is increasing steadily and some other factor covid/unprecedented factor can cause drastic changes in the expenditure. The conclusion revealed that mental health expenditure is steadily increasing, with unexpected factors like COVID-19 having the potential to cause significant fluctuations in this spending.

PREDICTION MODEL:

The goal of this analysis is to forecast changes in mental health expenditure for the Australian government, influenced by factors like drug use, unemployment, and social media. This is a regression problem, as it involves predicting a continuous variable—the government’s mental health expenditure—using multiple independent variables. The target variable in this study is the Australian government's mental health expenditure.

DATA PRE-PROCESSING:

Although the datasets were sourced from reputable sources and showed minimal discrepancies, there were a few NaN or null values that were imputed to preserve data integrity before model training. This step was crucial because each data point contributes to the analysis. Additionally, data preprocessing involved normalization to ensure consistency across different scales and addressing any missing values to avoid potential biases in the model. Feature selection was applied to identify the most relevant features affecting the target variable, reducing data noise through techniques like correlation heatmaps helping to identify and eliminate highly correlated or irrelevant features. Additionally, a forward feature selection method was employed, which iteratively added features to the model based on their contribution to predictive accuracy. This approach ensured that only the most significant variables were retained, thereby minimizing overfitting and enhancing the model's generalizability. This process not only lowered the dataset's dimensionality but also improved the model's performance and interpretability by focusing on the key factors influencing mental health expenditure.

MODEL SELECTION:

The selected models for this analysis encompass a range of 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 capitalize on their distinct strengths in addressing various characteristics of the data. Linear Regression offers a straightforward approach, while Ridge and Lasso Regression incorporate regularization to mitigate overfitting. ElasticNet combines the

penalties of both Ridge and Lasso, providing greater flexibility. Decision Trees and Random Forests are capable of capturing complex, non-linear relationships, while Gradient Boosting enhances predictive accuracy through iterative improvements. Support Vector Regression (SVR) further contributes by effectively modeling non-linear relationships in high-dimensional data. Collectively, these models enable a thorough evaluation of the factors influencing mental health expenditure and ensure the robustness of the predictions.

METHODOLOGY:

The methodology for this analysis comprised several essential steps to ensure robust model performance and accurate predictions. Initially, data was gathered from various reputable sources and integrated to form a comprehensive dataset containing relevant features and target variables. A preprocessing pipeline was established to handle data cleaning, normalization, and imputation, which helped maintain consistency and prevent data leakage by incorporating these steps into the model training process. The dataset was then partitioned into training and testing sets using a 70-30 split ratio to facilitate effective model training and evaluation. To improve model performance and reduce noise, irrelevant columns were removed, and a forward feature selection method was applied to identify the most significant features impacting the target variable. A diverse set of regression-based models was evaluated, including Linear Regression, Ridge Regression, Lasso Regression, ElasticNet, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Support Vector Regression (SVR). Hyperparameters for each model were optimized using GridSearchCV, which systematically explored different combinations to identify the best settings. The models were then assessed using the R^2 metric to measure the proportion of variance explained. The top-performing models were selected based on their R^2 scores, ensuring that 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

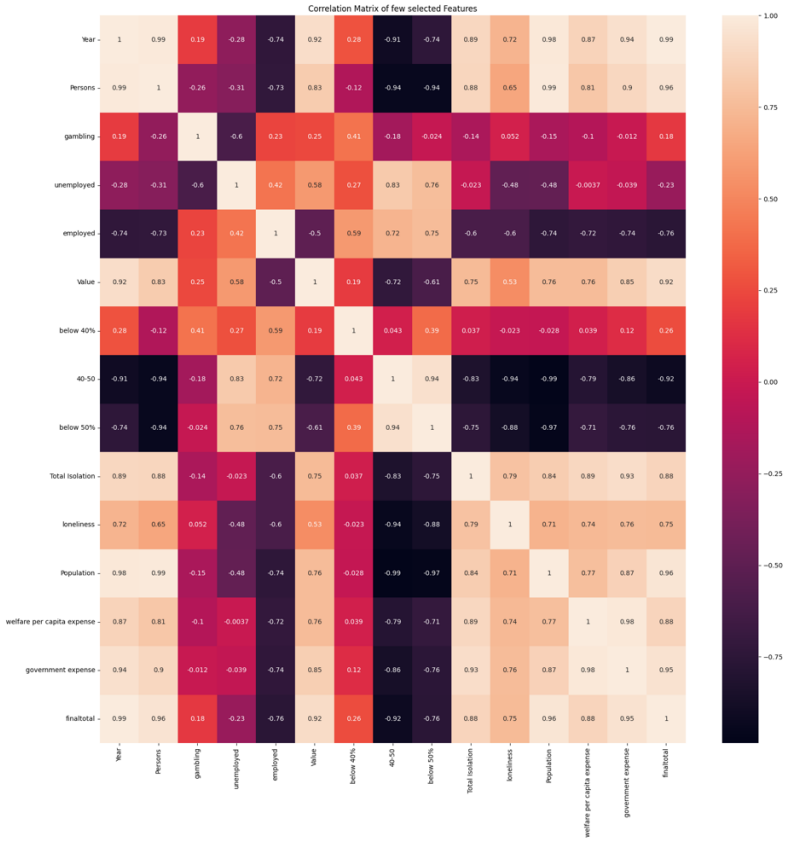

Figure 2 Heat Map for selected features from the dataset

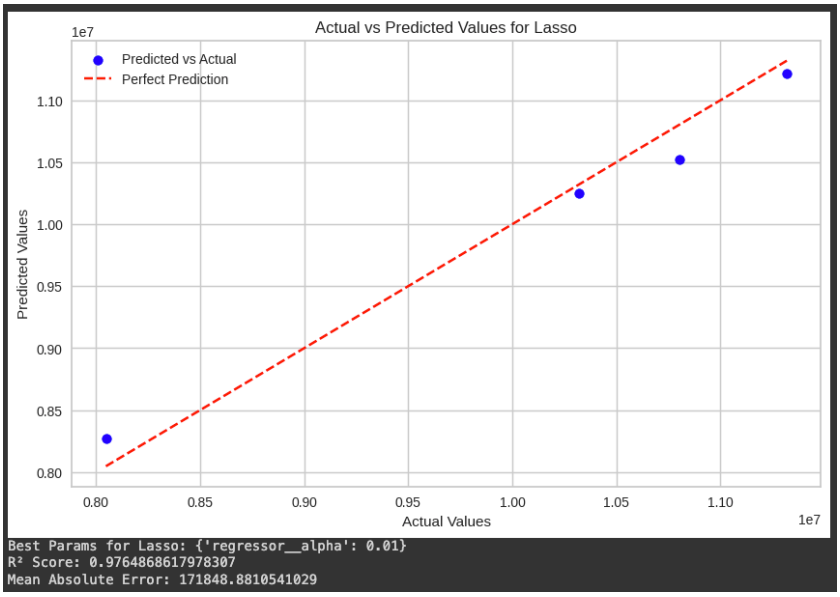

Figure 3 Model's performance on Test Set

FUTURE WORK:

Based on the analysis, several specific actions can be recommended to enhance the model's performance and reliability. Given the limited dataset, acquiring additional data or using synthetic data generation techniques could significantly improve model accuracy and generalization. Enhance feature engineering by conducting a thorough feature importance analysis and incorporating domain knowledge to refine or create new features. Expand the hyperparameter search space using RandomizedSearchCV for a more efficient exploration of optimal settings. Apply k-Fold Cross-Validation and continue using Leave-One-Out Cross-Validation to better estimate model performance and reduce variability. Finally, prepare the best-performing model for deployment, ensuring it integrates seamlessly into existing systems, and set up continuous monitoring with a feedback loop to track performance and make iterative improvements.

RESULT AND CONCLUSION:

The results demonstrate that the Linear Regression model with Lasso regularization has proven to be highly effective, achieving an impressive R^2 score of 0.97, which indicates that it explains 97% of the variance in mental health expenditure. This high level of accuracy reflects the model's strong fit and predictive power. Lasso regularization not only improved performance by addressing overfitting but also streamlined the model by selecting the most significant features, enhancing its interpretability. This model offers valuable insights into the impact of key factors such as drug use, unemployment, and social media on mental health expenditure. Its robust performance makes it a powerful tool for policymakers and stakeholders, enabling informed decisions on resource allocation and strategic planning. However, continuous evaluation and testing with additional data are recommended to ensure its scalability and generalizability in various contexts. The government's spending on mental health is expected to rise, and new factors will likely emerge that influence an individual's mental well-being. Although the analysis was successful, there's potential to delve deeper into the factors affecting mental health, which could further enhance the model's accuracy and insights.

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.