# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Shivani Nandkishor Nipane |
| **Project Name** | ASSESSMENT TASK1: Regression Models |
| **Date** | 03/04/2023 |
| **Deliverables** |  |

|  |  |
| --- | --- |
| 1. **EXPERIMENT BACKGROUND** | |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | |
| **1.a. Business Objective** | Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?   * This project aims to accurately predict cancer mortality rates based on information related to US counties using a dataset consolidated from census data. The dataset contains 33 different features, including demographic and medical information. The business aims to identify patterns, correlations, and factors contributing to cancer mortality rates in different regions by developing a predictive model. * The results of this project can be used by healthcare organizations, government agencies, and policymakers to allocate resources more efficiently, plan targeted interventions, and develop tailored cancer prevention and control programs. By understanding which factors significantly impact cancer mortality rates, decision-makers can prioritize specific actions and strategies to address the most critical issues and reduce the overall cancer burden. * The impact of accurate results is significant. Accurate predictions can lead to better-informed decisions, more effective allocation of resources, improved healthcare services, and, ultimately, reduced cancer mortality rates. Accurate results can also help identify disparities in healthcare access and outcomes, enabling decision-makers to address inequalities and improve the population's overall health. * On the other hand, incorrect results may lead to the misallocation of resources, ineffective interventions, and misguided policies, which can further exacerbate existing healthcare disparities and contribute to higher cancer mortality rates. Inaccurate predictions may also result in a loss of trust in the model, undermining its potential benefits and limiting its adoption by stakeholders. * Therefore, it is crucial to validate the model's performance using appropriate evaluation metrics, such as mean squared error (MSE) and R-squared score. Iteratively refine the model to ensure its accuracy and reliability. Additionally, the model should be periodically updated and reassessed as new data becomes available to maintain its relevance and accuracy. |
| **1.b. Hypothesis** | Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it,   * **Hypothesis:** Certain demographic and medical factors can significantly predict cancer mortality rates in US counties. * The question we want to answer is: Which factors strongly influence cancer mortality rates, and how can we use this information to develop a predictive model that can accurately estimate cancer mortality rates for different US counties? * We are seeking to understand the relationships between demographic and medical factors and their impact on cancer mortality rates, enabling healthcare organizations, government agencies, and policymakers to make better-informed decisions and implement targeted interventions to reduce cancer mortality rates.   This hypothesis is worth considering for several reasons:   1. Identifying key factors: By investigating the relationships between various demographic and medical factors and cancer mortality rates, we can identify the most important factors contributing to the variation in cancer mortality rates across US counties. This information can help prioritize targeted interventions and resources to address the most significant factors. 2. Resource allocation: An accurate predictive model that can estimate cancer mortality rates based on demographic and medical factors will enable healthcare organizations and government agencies to allocate resources more effectively, focusing on regions with higher predicted cancer mortality rates and the factors contributing to these rates. 3. Targeted interventions: Understanding the factors contributing to cancer mortality rates can help develop tailored cancer prevention and control programs. It may include targeted awareness campaigns, healthcare infrastructure improvements, early detection and screening programs, and treatment accessibility to address each county's specific needs and challenges. 4. Health disparities: Exploring the factors influencing cancer mortality rates can reveal underlying health disparities and inequalities in healthcare access, quality, and outcomes. By addressing these disparities, policymakers can work towards ensuring equitable healthcare services for all populations, regardless of their demographic or medical characteristics. 5. Monitoring and evaluation: By analyzing the factors that influence cancer mortality rates and their relative importance, we can establish a framework for monitoring and evaluating the effectiveness of cancer control policies and programs, enabling continuous improvement and evidence-based decision-making. |
| **1.c. Experiment Objective** | Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.   * The expected outcome of this experiment is to develop a predictive model that accurately estimates cancer mortality rates in US counties based on demographic and medical factors. We expect to identify the most influential factors contributing to cancer mortality rates and quantify their impact. * We aim to achieve an R-squared score of at least 0.7, indicating that the model can explain 70% or more of the variance in cancer mortality rates. This threshold suggests a reasonably accurate model that can be used for decision-making.   Possible scenarios resulting from this experiment:   1. Strong predictive model: We develop a model that meets or exceeds our goal with an R-squared score of 0.7 or higher. In this case, the model can be considered reliable for predicting cancer mortality rates and identifying key factors. The model can be used by healthcare organizations, government agencies, and policymakers to inform their decisions, allocate resources, and develop targeted interventions. 2. Moderate predictive model: The model achieves an R-squared score between 0.5 and 0.7, suggesting moderate predictive power. While not as reliable as a strong model, it can still provide valuable insights into the relationships between demographic and medical factors and cancer mortality rates. Further refinement, additional data, or more advanced modelling techniques may be needed to improve the model's accuracy. 3. Weak predictive model: The model has an R-squared score below 0.5, indicating limited predictive power. In this scenario, the model may not be suitable for decision-making or resource allocation. We need to reevaluate the features, model selection, or data quality and explore alternative approaches to improve the model's performance. 4. No significant relationships: The model may fail to identify significant relationships between demographic and medical factors and cancer mortality rates. In this case, we would need to reconsider the features used in the model, explore alternative data sources, or investigate other factors that may influence cancer mortality rates beyond the scope of the current dataset. 5. Model overfitting or underfitting: The model may overfit or underfit the data, leading to poor performance on new data. In this case, we need to revisit the model's complexity, regularization parameters, or feature selection techniques to improve generalization and prevent overfitting or underfitting. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS** | |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | |
| **2.a. Data Preparation** | Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments.  Steps taken for preparing the data:   1. Data cleaning: The initial step involved handling missing or inconsistent values, such as imputing missing values using appropriate techniques (e.g., mean, median, or mode imputation) or removing rows with a high percentage of missing data. This step ensures the dataset is complete and consistent, improving the model's performance and reliability.   **Rationale:** Incomplete or inconsistent data can lead to inaccurate predictions and may introduce bias into the model.   1. Exploratory Data Analysis (EDA): We conducted an EDA to understand the data's distribution, identify outliers, and detect potential relationships between features and the target variable. This step included generating summary statistics, visualizations (e.g., histograms, box plots, scatter plots), and correlation matrices.   **Rationale:** EDA provides a deeper understanding of the data, helps identify potential issues, and informs feature selection and model choice.   1. Feature selection: Based on the insights from EDA and correlation analysis, we selected relevant features that exhibited a strong relationship with the target variable, cancer mortality rates.   **Rationale:** Selecting relevant features improves model performance, reduces overfitting, and simplifies interpretation.   1. Data transformation: We applied data transformations, such as normalization or standardization, to ensure that features are on a comparable scale. This step is particularly important for models sensitive to feature scaling, such as linear regression and distance-based algorithms.   **Rationale:** Data transformation ensures the model is not biased towards larger-scale features and improves model convergence during optimization.  Steps not executed and reasoning:   1. Feature engineering: Although we did not create additional features in this experiment, future experiments could benefit from exploring interactions between existing features, creating polynomial features, or applying domain knowledge to generate new features.   **Reasoning:** We focused on identifying the most influential factors in the existing dataset and simplifying the model for interpretability.  Important steps for future experiments:   1. Feature engineering: Investigating interactions between features or creating new features based on domain knowledge may reveal additional insights and improve model performance. 2. Model selection: Exploring alternative models, such as ensemble methods (e.g., Random Forest, Gradient Boosting) or more advanced techniques (e.g., Neural Networks), could improve the predictive accuracy and robustness. 3. Hyperparameter tuning: Optimizing model hyperparameters using techniques such as Grid Search or Random Search can help achieve better performance and prevent overfitting or underfitting. 4. Cross-validation: Implementing cross-validation techniques, such as k-fold cross-validation, can provide a more reliable estimation of the model's performance on unseen data and reduce the risk of overfitting. 5. Regularization: Incorporating regularization techniques, such as Lasso, Ridge, or Elastic Net, can help prevent overfitting and improve the model's generalization to new data |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments.   1. Feature selection: Based on the insights from EDA and correlation analysis, we selected relevant features that exhibited a strong relationship with the target variable, TARGET\_deathRate.   **Rationale:** Selecting relevant features improves model performance, reduces overfitting, and simplifies interpretation.  This step was essential to find which feature has more correlation with the target variable, there was a strong correlation between incidenceRate,povertyPercent and PctPublicCoverageAlone.  So I decided to use this use these features for better performance of the model.  I removed some of the features such as PctSomeCol18\_24 as it had more null values and i also dropped Geography column as it as more only one count in each unique value  and dropping the binnedInc column as we already have Medianincome of a country column. |
| **2.c. Modelling** | Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments   1. Univariate Linear Regression: We chose univariate linear regression, a simple and interpretable model that captures the relationship between a single feature and the target variable. This model provides an initial baseline performance and helps us understand the importance of the selected feature in predicting the target variable. We did not need to tune any hyperparameters in this model. 2. Multivariate Linear Regression: We chose this model to investigate the relationship between multiple features and the target variable. This approach enables us to capture more complex relationships in the data and achieve better performance. Like univariate linear regression, there are no hyperparameters to tune for multivariate linear regression.   We trained a multivariate linear regression model with feature engineering for this experiment. The reasons for choosing this model are as follows:   1. Multivariate linear regression can handle multiple features, which allows us to capture complex relationships between the features and the target variable. 2. Linear regression models are simple, interpretable, and can perform well when the relationships between features and the target variable are linear or close to linear. 3. Feature engineering can help improve model performance by transforming the raw data into a more suitable format, such as scaling or applying log transformation to the features.   There are no hyperparameters to tune for a linear regression model. But we used techniques from feature engineering, such as log transformation and standardization, to make the model work better. The rationale for using these techniques is as follows:   1. Log transformation: This can help reduce the influence of outliers and transform skewed distributions into more symmetric ones. This can lead to better performance for linear regression models. 2. Standardization: Scaling the features with a mean of 0 and a standard deviation of 1 can help the model converge faster and improve performance.   We decided not to train the following models:   1. Decision Trees, Random Forests, and Gradient Boosting: These models are more complex than linear regression models and may not be necessary for this problem. However, if the linear regression models do not provide satisfactory results, these models can be considered for future experiments. 2. Support Vector Machines: SVMs can be computationally expensive, and linear regression models perform better with less computational resources. If linear regression models are insufficient, SVMs with different kernels could be considered for future experiments.   For future experiments, it might be worthwhile to investigate more complex models, such as ensemble methods or deep learning models, to improve performance further. More advanced feature engineering techniques and feature selection methods could also be explored. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | |
| **3.a. Technical Performance** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.    Unvariate Linear Regression:  The two features which we selected for unvariate linear regression are incidenceRate and povertyPercent.  The relevant performance of these two features are as follows:   * Feature 1: IncidenceRate:   Mean Squared Error: 621.3817645016619  R-squared Score: 0.23906669796187174   * The performance of the model is less than 0.5 as there no advance algorithms used to train this model. If we used different model its performance can increase. * Feature 2: povertyPercent:   Mean Squared Error: 693.7361519543474  R-squared Score: 0.1504627734397661   * The performance of the model is less than 0.5 as there no advance algorithms used to train this model. If we used different model its performance can increase.   Multivariate Regression Model:  In this we trained all the numeric columns present in the dataset.  The relevant performance of the multivariate regression model as follows:  Mean Squared Error: 422.95313238675396  R-squared Score: 0.4820589498751512   * The performance of the model is close to 0.5 as there no advance algorithms used to train this model. If we used different model its performance can increase.   So we trained the model using feature engineering  Mean Squared Error: 259.905417623331  R-squared: 0.5674975513002033  The performance of the model is between 0.5 to 0.7 which means the performance of the model has been increased using the feature engineering which is moderate performace. |
| **3.b. Business Impact** | Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)  Interpreting the results of the experiments related to the business objective of predicting cancer mortality rates in US counties can be done by analyzing the model's performance, such as the Mean Squared Error (MSE) and R-squared score. A lower MSE and a higher R-squared score indicate better performance, meaning the model can better predict cancer mortality rates based on the provided features.  The impacts of incorrect results for the business can vary depending on the specific use case:   1. Resource allocation: If the model's predictions are used to allocate resources for cancer treatment and prevention, incorrect results may lead to an inefficient allocation of resources. Overestimating cancer mortality rates in certain areas may cause an excessive concentration of resources, while underestimating rates might leave some areas underprepared to deal with the cancer burden. 2. Public health policies: Inaccurate predictions can influence public health policies and initiatives, possibly leading to the implementation of ineffective strategies or the misallocation of funds. This could ultimately hinder the progress towards reducing cancer mortality rates. 3. Healthcare planning: Healthcare providers may use these predictions to plan their services and facilities, such as the number of cancer treatment centres or the distribution of specialized medical professionals. Incorrect results may cause inadequate planning, leading to potential shortages or surpluses of resources in certain areas. 4. Research priorities: The model's predictions could inform the prioritization of research efforts to identify factors contributing to high cancer mortality rates in specific regions. If the model provides accurate results, research priorities may be correctly guided, and essential areas might not receive attention.   To mitigate the potential negative impacts of incorrect results, it is essential to continuously validate and improve the model with updated data, explore different models and feature engineering techniques, and consider incorporating domain expertise into the decision-making process. Additionally, it is crucial to maintain transparency about the model's limitations and uncertainties when communicating results to stakeholders. |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.  During the experiments, several issues were faced, both solved and unsolved:   1. Missing or incomplete data: Some records in the dataset had missing or incomplete values, which could affect the model's performance. To address this issue, we used imputation techniques to fill in the missing values, such as using the mean or median of the respective feature. Gathering more complete data or investigating more advanced imputation methods could be beneficial for future experiments. 2. Feature selection: Choosing the most relevant features for the model was challenging. We used correlation analysis to identify features strongly related to the target variable. However, this method might only capture some relevant features or the best combination of features. Other feature selection techniques, such as Recursive Feature Elimination (RFE) or Lasso regularization, could be considered for future experiments. 3. Non-linear relationships: Linear regression models assume a linear relationship between features and the target variable. If some relationships are non-linear, the model's performance may be limited. We applied log transformation and standardization during the feature engineering step to address this issue. Exploring non-linear models or advanced transformation techniques may help improve performance for future experiments. 4. Model selection: While we focused on linear regression models for this experiment, other models may perform better. Due to their higher complexity and computational cost, we did not train more complex models such as decision trees, random forests, or support vector machines. However, it could be valuable to explore these models in future experiments if linear regression models do not provide satisfactory results. 5. Hyperparameter tuning: Although linear regression models do not have hyperparameters to tune, other models like ElasticNet or Support Vector Machines do. Exhaustive hyperparameter tuning can be computationally expensive and time-consuming. In this experiment, we used default hyperparameters or basic grid search techniques. For future experiments, more advanced hyperparameter tunings methods like Bayesian optimization or randomized search could be used to optimize the model's performance. 6. Overfitting: Complex models may overfit the training data, leading to a poor generalization of unseen data. To mitigate overfitting, we used regularization techniques like ElasticNet. Applying other regularization methods or cross-validation techniques for future experiments can help prevent overfitting and improve model performance.   If these problems are fixed in future experiments, the models will work better and predictions will be more accurate, giving the business more value. |

|  |  |
| --- | --- |
| 1. **FUTURE EXPERIMENT** | |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | |
| **4.a. Key Learning** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.  Reflecting on the outcome of the experiment, we gained several insights:   1. Feature selection is crucial: We observed that selecting the most relevant features significantly impacted the model's performance. This highlights the importance of proper feature selection and engineering techniques, which can lead to better models and improved predictions. 2. Linear regression models can provide valuable insights: Although linear regression models have limitations, they can still provide valuable information about the relationship between features and the target variable. They serve as a good starting point for more complex models and can help understand the underlying patterns in the data. 3. Data quality matters: The quality of the data, including the presence of missing or incomplete values, can directly impact the performance of the models. Ensuring the data is clean and well-prepared is essential for building reliable models. 4. Regularization techniques can help mitigate overfitting: Using techniques like ElasticNet helped prevent overfitting and improved the generalization of the models on unseen data.   Based on these new ideas, it makes sense to try out the current method more, since there is still room for improvement. However, to overcome some of the limitations of linear regression models, exploring other machine learning algorithms, such as decision trees, random forests, or support vector machines may be valuable. These models can capture non-linear relationships in the data and may provide better performance.  Additionally, incorporating domain expertise and expanding the feature set may lead to better models. More advanced feature engineerings techniques, such as interaction terms or polynomial features, can also be explored to enhance the model's predictive power.  In short, the current method has given us some good information and a good starting point for predicting cancer death rates, but there is still room for improvement. More testing with different models, feature engineering techniques, and data preparation methods can improve performance and make predictions that are more accurate. |
| **4.b. Suggestions / Recommendations** | Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production.  Based on the results and the overall goal of the project, the following are some possible next steps and experiments:   1. Explore other machine learning algorithms:    1. Decision Trees    2. Random Forests    3. Support Vector Machines    4. Gradient Boosting Machines    5. Neural Networks Expected uplift: High Rationale: These models can capture non-linear relationships and complex interactions between features, potentially leading to better performance. 2. Advanced feature engineering techniques:    1. Interaction terms    2. Polynomial features    3. Domain-specific feature generation Expected uplift: Medium Rationale: Enhancing the feature set may improve the model's predictive power and provide better insights into the underlying patterns in the data. 3. Improved hyperparameter tuning:    1. Bayesian optimization    2. Randomized search    3. Genetic algorithms Expected uplift: Medium Rationale: Optimizing hyperparameters can improve performance by fine-tuning the models to the specific dataset. 4. Incorporate domain expertise:    1. Collaborate with domain experts to identify potential relevant features or insights. Expected uplift: Medium Rationale: Leveraging domain expertise can lead to a better understanding of the data and the relationships between features and the target variable, potentially improving model performance. 5. Ensemble models:    1. Combine multiple models to create a more robust predictive model. Expected uplift: Medium Rationale: Ensemble models can leverage the strengths of different algorithms and reduce the impact of individual model weaknesses, improving overall performance. |