Medical Expenditure Analysis Report

Introduction:

Our final model, a Random Forest classifier, has shown exceptional predictive performance in identifying individuals with large expenditures. The Random Forest algorithm, an ensemble learning method, constructs multiple decision trees during training and outputs the mode of the classes for classification.

Steps Taken:

1. Business Understanding and Data Understanding:

- Explored the business context and objectives.

- Loaded the dataset and acquired information on column data types.

- Conducted Exploratory Data Analysis (EDA) for a comprehensive understanding of the data.

2. Data Preparation:

- Checked and handled missing values.

- Ensured consistency by identifying and addressing duplicates.

- Verified uniformity in column data types.

- Validated data by identifying and handling outliers.

- Created the target column indicating expenditures greater than $20,000 as 1 and less as 0.

- Employed label encoding for categorical columns.

- Scaled the data to standardize features.

- Handled class imbalance using Synthetic Minority Over-sampling Technique (SMOTE).

- Split the data into a training set (75%) and a test set (25%).

3. Modeling:

- Developed baseline Logistic Regression model.

- Implemented a Decision Tree model.

- Utilized the Random Forest model, which emerged as the final model due to superior performance.

4. Model Evaluation:

- Achieved a remarkable 97% cross-validation score.

Answers to Questions:

1. Sensitivity and Specificity:

- Calculated sensitivity and specificity, both achieving a score of 1.0, indicating excellent performance.

2. Comparison with Simpler Model:

- Compared the final model with a simpler model, where the simpler model also achieved a 97% cross-validation score.

3. AUC Analysis:

- AUC of the simpler model, with specific features, demonstrated a score of 97% without cross-validation and a perfect score of 1.0 with validation. This highlights the significant impact of cross-validation on model performance.

4. Summary of the Final Model:

- The Random Forest model performed exceptionally well, with perfect cross-validation scores and a flawless confusion matrix. However, caution is warranted due to the observed class imbalance in the target column. Further analysis is recommended, involving gathering more data and exploring additional feature engineering techniques to enhance model robustness and generalizability.

Recommendations for Future Work:

- Collect additional data to address class imbalance.

- Explore more advanced feature engineering techniques.

- Conduct a thorough analysis of model interpretability and potential biases