KARNAVATI UNIVERSITY

UNITEDWORLD SCHOOL OF COMPUTATIONAL INTELLIGENCE (USCI)

Summative Assessment (SA)

Submitted by
Tanya Chhabhadiya
(Enrl. No.: 20210701007)

Course Code and Title: 21BSCS35E06 - Predictive Analytics

B.Sc. (Hons.) Computer Science / Data Science / AIML
V Semester – July – Nov 2023

USCI

Nov/Dec 2023

TABLE OF CONTENTS

1. ABSTRACT

In order to handle the rising costs of healthcare and give people and families a financial safety net against unforeseen medical expenses, medical insurance pricing is necessary. People could have trouble managing the cost of healthcare and getting access to essential medical services if they don't have enough insurance. Age, lifestyle choices, pre-existing diseases, and regional differences in healthcare expenses are just a few of the variables that are taken into account when pricing medical insurance. These elements are reflected in actuarial calculations that guarantee the sustainability of insurance policies.

The project on Medical Insurance Price Prediction leverages predictive analytics to estimate health insurance charges based on various demographic and health-related features. The dataset used for analysis encompasses information such as age, sex, BMI, number of children, smoking habits, and region. The project involves a comprehensive data preprocessing phase, including handling missing values, visualizing feature distributions, handling outliers, and encoding categorical variables. Exploratory data analysis is performed to gain insights into the relationships between different features and the target variable—insurance charges.

The data modeling phase involves building and evaluating predictive models to accurately estimate insurance charges. Several regression models, including Linear Regression, Random Forest Regression, Gradient Boosting Regression, and XGBoost Regression, are employed. Hyperparameter tuning is performed to optimize the models, enhancing their predictive capabilities. Feature importance analysis is conducted to identify the key factors influencing insurance charges.

The final model, an XGBoost Regressor, is chosen based on its superior performance. The project achieves a high level of accuracy in predicting insurance charges, as evidenced by the evaluation metrics on both training and testing datasets. The predictive model provides a valuable tool for insurance providers to estimate charges based on individual characteristics, facilitating informed decision-making and personalized insurance offerings. This project exemplifies the application of predictive analytics in the healthcare domain, contributing to data-driven decision-making and improved resource allocation.

Keywords: Data Mining, Medical Insurance, Predictive Modeling, Regression Algorithms, Exploratory Data Analysis, Hyperparameter Tuning, Feature Importance, XGBoost, Data Preprocessing, Outlier Handling, Cross-Validation.

2. INTRODUCTION

Medical insurance acts as a safety net, shielding consumers' finances from unforeseen, frequently high medical expenses. People may have financial difficulties as a result of inadequate health insurance, which may force them to make difficult decisions between their health and financial security.

The "Medical Insurance Price Prediction" project is a pioneering endeavor in the realm of predictive analytics, designed to revolutionize the estimation of health insurance charges. In an era where data-driven insights play a pivotal role in decision-making, this project harnesses the power of advanced analytics to predict insurance costs based on a diverse set of demographic and health-related factors. The dataset employed for this project encapsulates crucial information, including age, gender, BMI, number of children, smoking habits, and region.

The journey begins with an extensive data preprocessing phase, encompassing tasks such as handling missing values, outlier detection, and encoding categorical variables. Visualizations are employed to unravel the intricate relationships between different features and insurance charges. The subsequent data modeling phase involves the deployment of various regression models, including Linear Regression, Random Forest Regression, Gradient Boosting Regression, and XGBoost Regression.

In the pursuit of model optimization, hyperparameter tuning is executed to fine-tune the algorithms for enhanced predictive accuracy. The identification of key features influencing insurance charges through feature importance analysis adds depth to the project's insights. The final model, an XGBoost Regressor, emerges as the epitome of predictive prowess, delivering superior performance in estimating insurance charges.

This project not only exemplifies the convergence of healthcare and data science but also underscores the potential for informed decision-making and personalized insurance offerings. As we delve into the intricacies of predictive analytics, the "Medical Insurance Price Prediction" project stands as a beacon of innovation, contributing to the evolving landscape of data-driven healthcare solutions.

3. RELATED WORKS

Several studies in the field of medical insurance pricing and predictive modeling have contributed valuable insights, shaping the landscape of healthcare financing. Prior research has often focused on exploring the factors influencing insurance premiums, understanding the dynamics of healthcare costs, and improving the accuracy of predictive models. One notable study, for instance, delved into the impact of lifestyle choices, including smoking habits, on insurance pricing. The findings highlighted the substantial effect of smoking on healthcare costs, influencing both individual health and the broader financial landscape of insurance providers.

Additionally, studies have examined the role of demographic factors, such as age and gender, in shaping insurance premiums. The implications of an aging population, coupled with the increasing prevalence of chronic conditions, have been a focal point in understanding the evolving risk profiles within insured populations. These insights have paved the way for more nuanced and tailored pricing structures that better align with the diverse health needs of different demographic groups.

The use of advanced predictive modeling techniques, akin to the approach taken in this project, has also been a significant area of exploration. Studies employing algorithms like Random Forest, Gradient Boosting, and XGBoost have demonstrated enhanced predictive accuracy in forecasting insurance charges. These models have shown promise in capturing complex relationships among various predictors, offering more robust and reliable predictions compared to traditional linear models.

The impact of these related works has been substantial in reshaping insurance pricing strategies and fostering a data-driven approach in the healthcare financing domain. By incorporating insights from diverse studies, the project at hand builds upon this foundation, aiming to contribute additional depth and specificity to the predictive modeling landscape. The utilization of outlier handling techniques, feature importance analysis, and hyperparameter tuning in this project reflects a progressive integration of methodologies inspired by previous works. As a result, the project not only advances the understanding of medical insurance pricing but also showcases the practical implications of incorporating diverse data mining techniques for optimizing predictive performance in this critical field.

4. RESEARCH GAPS IDENTIFIED

While the project on medical insurance price prediction has made significant strides in leveraging data mining techniques, several research gaps and opportunities for further exploration have been identified within this domain. One notable gap lies in the limited consideration of temporal dynamics in the dataset. The existing project primarily focuses on static features such as age, BMI, and smoking status, but fails to incorporate time-dependent variables that may play a crucial role in understanding the evolving nature of healthcare costs. Future research could explore the inclusion of time-series data, allowing for a more dynamic and nuanced analysis of how insurance charges change over time.

Another research gap pertains to the potential influence of external factors, such as economic indicators or healthcare policy changes, on medical insurance pricing. The existing project primarily examines individual-level factors, but broader contextual variables might contribute significantly to variations in healthcare costs. Exploring the integration of external datasets that capture economic trends or policy shifts could provide a more comprehensive understanding of the determinants of insurance charges.

Furthermore, while the project considers demographic and lifestyle factors, there is an opportunity to delve deeper into the intersectionality of these variables. Research that examines how combinations of demographic attributes and lifestyle choices interact to impact insurance pricing could offer more nuanced insights. For instance, understanding how the interaction between age and smoking status affects insurance charges could lead to more targeted and tailored pricing models.

The project also identifies the need for more extensive exploration into feature engineering techniques. While the initial project preprocesses the data, incorporating advanced feature engineering methodologies could enhance the predictive power of models. Techniques such as interaction terms, polynomial features, or domain-specific transformations may reveal hidden patterns and relationships within the data, contributing to more accurate predictions.

In summary, the research gaps identified in this project highlight the potential for further advancements in the application of data mining to medical insurance pricing. Future studies addressing these gaps could lead to a more holistic understanding of the multifaceted factors influencing insurance charges, thereby refining predictive models and contributing to the ongoing evolution of healthcare financing strategies.

5. PROPOSED METHODOLOGY

The methodology for the medical insurance price prediction project follows a systematic approach, beginning with the loading of the dataset using the Pandas library. This initial step involves gaining an understanding of the dataset's structure, variable types, and overall scale, setting the foundation for subsequent analyses.

Data cleaning and preprocessing are crucial components of the methodology. Using Pandas, missing values are addressed to ensure data completeness, and duplicate records are removed to enhance the dataset's reliability. The Feature Engine library is employed for outlier handling, particularly in the 'bmi' variable, where extreme values are capped to promote a more robust dataset for analysis.

The exploratory data analysis (EDA) phase utilizes Matplotlib and Seaborn for visualization. Pie charts illustrate the distribution of categorical variables such as 'sex,' 'smoker,' and 'region.' Bar plots provide insights into the mean insurance charges across different categories, while scatter plots explore relationships between numerical features ('age' and 'bmi') and insurance charges. The handling of outliers is a critical step, involving the Feature Engine library to cap extreme values in the 'bmi' variable. The impact of this outlier handling is visually represented through box plots, showcasing the reduction in extreme values. A correlation heatmap, generated using Seaborn, visually represents relationships among numerical variables, aiding in the identification of potential multicollinearity and understanding variable correlations.

Feature importance analysis, conducted using the XGBoost Regressor, guides the selection of influential features for the final predictive model. The importance scores are visualized through a DataFrame. The overall visual appeal of the project is emphasized, making complex patterns and relationships accessible to both technical and non-technical stakeholders through visualization tools. The data mining perspective comes to the forefront during the predictive modeling phase, where diverse regression algorithms are implemented, including Linear Regression, Random Forest Regression, Gradient Boosting Regressor, and XGBoost Regression. The Scikit-learn library facilitates model implementation and evaluation, while hyperparameter tuning using GridSearchCV optimizes model parameters. The final model is refined by eliminating non-contributing features based on feature importance analysis.

The project's execution takes place in Jupyter Notebooks and Google Colab, providing an efficient environment for code development and execution. The methodology, rooted in data mining principles, ensures a systematic and exploratory approach to uncovering patterns, ensuring data quality, and optimizing predictive models for medical insurance pricing.

6. DATASET DESCRIPTION

The dataset used for the medical insurance price prediction project consists of several columns, each providing information about specific aspects of individuals and their medical insurance characteristics. Here is a description of each column in the dataset:

1. **age:** Represents the age of the insured individual. This variable provides insight into how age may influence medical insurance pricing.
2. **sex:** Indicates the gender of the insured individual, with 'male' and 'female' as possible values. This variable helps assess gender-based differences in insurance charges.
3. **bmi:** Stands for Body Mass Index, a measure derived from an individual's weight and height. BMI is relevant in understanding the impact of body composition on insurance charges.
4. **children:** Denotes the number of children or dependents covered by the insurance plan. This variable considers the family structure and its potential influence on insurance pricing.
5. **smoker:** Indicates whether the insured individual is a smoker or not, with 'yes' and 'no' as possible values. Smoking status is a significant factor influencing healthcare costs and, consequently, insurance charges.
6. **region:** Represents the geographical region of the insured individual. The dataset includes four regions: 'northwest,' 'northeast,' 'southeast,' and 'southwest.' Regional variations may impact healthcare costs and insurance pricing.
7. **charges:** This is the target variable, representing the medical insurance charges incurred by the insured individual. The goal of the project is to predict this variable based on the other features in the dataset.

7. DATA PREPROCESSING

The data preprocessing phase of the "Medical Insurance Price Prediction" project is a crucial step in ensuring the reliability and effectiveness of the predictive models. The dataset is first loaded and its structure is explored using pandas, revealing the initial insights into the data. Exploratory data analysis involves checking for missing values, visualizing feature distributions, and gaining an understanding of the data's statistical properties. Duplicates are removed from the dataset to enhance data integrity.

Visualizations, including pie charts and bar plots, provide a comprehensive overview of categorical features such as sex, smoker status, and region, shedding light on their distributions. Scatter plots illustrate the relationships between numerical features like age and BMI with insurance charges, particularly highlighting the impact of smoking status. Outliers are detected and handled in the BMI feature using box plots and an arbitrary outlier capper.

Correlation heatmaps reveal the interplay between different features, informing decisions on feature selection and model building. Skewness in numerical features is addressed through appropriate transformations. Categorical variables undergo encoding for compatibility with machine learning algorithms. Overall, the data preprocessing phase ensures the dataset's cleanliness, handles outliers, and prepares the features for the subsequent model-building stage, laying a solid foundation for accurate medical insurance price predictions.

8. MODEL BUILDING

Linear Regression:

- **Description:** Linear Regression is a simple and widely-used supervised learning algorithm for predicting a continuous outcome. It assumes a linear relationship between the input features and the target variable. The model computes a linear equation by adjusting coefficients to minimize the difference between predicted and actual values.
- **Working:** The algorithm aims to find the best-fitting line through the data points by minimizing the sum of squared differences (residuals) between predicted and actual values. The coefficients are determined using the least squares method, making the model interpretable and easy to implement.

Implementation:

The Linear Regression model is then instantiated and trained on the training data, consisting of features (X) and the target variable (Y). Predictions are generated for the test set, and the model's performance is evaluated using the coefficient of determination (R-squared). The R-squared values indicate the proportion of the variance in the insurance charges explained by the model for both the training and test sets. Additionally, a cross-validation score is computed using 5-fold cross-validation to assess the model's generalization ability. These metrics collectively serve as a preliminary assessment of the Linear Regression model's accuracy and its capability to predict medical insurance charges.

Random Forest Regression:

- **Description:** Random Forest Regression is an ensemble learning method that combines multiple decision tree models to improve predictive performance. Each tree is built on a random subset of the dataset, and predictions are averaged or voted upon for the final result.
- **Working:** Decision trees are constructed based on feature importance, and the random forest aggregates their predictions to reduce overfitting and enhance generalization. It is robust against outliers and can capture complex relationships in the data. Hyperparameter tuning, such as adjusting the number of trees (n_estimators), helps optimize model performance.

Implementation:

In the subsequent phase of model building, the Random Forest Regressor is implemented for predicting medical insurance charges. Initially, the model is constructed with default parameters, and predictions are made on both the training

and test sets. The evaluation metrics, including R-squared for both training and test sets, along with the cross-validation score, are printed to assess the initial performance "Before Tuning."

To enhance the model's effectiveness, hyperparameter tuning is conducted using GridSearchCV. The optimal number of estimators is determined as part of the tuning process. The model is then re-instantiated with the optimized hyperparameters, and predictions are recalculated. The refined model, "After Tuning," showcases improved performance, demonstrated by higher R-squared values and a more reliable cross-validation score. This iterative approach, involving initial model construction, hyperparameter tuning, and subsequent evaluation, ensures the Random Forest Regressor is fine-tuned to better predict medical insurance charges.

Gradient Boosting Regression:

- **Description:** Gradient Boosting Regression is another ensemble method that builds a series of weak learners (usually decision trees) sequentially, with each tree correcting the errors of its predecessor. It combines their predictions to create a powerful ensemble model.
- **Working:** The algorithm minimizes a loss function by fitting a tree to the residuals of the previous one. This iterative process continues, with each tree addressing the model's weaknesses. Gradient boosting is effective but may require careful tuning to prevent overfitting.

Implementation:

In the model development phase, the Gradient Boosting Regressor is implemented to predict medical insurance charges. Initially, the model is created with default parameters, and predictions are generated for both the training and test sets. Evaluation metrics, including R-squared values for both sets and the cross-validation score, are printed to assess the initial performance "Before Tuning."

To optimize the model's performance, hyperparameter tuning is conducted using GridSearchCV. The optimal values for the number of estimators and learning rate are identified during this process. The model is then re-instantiated with the tuned hyperparameters, and predictions are recalculated. The refined model, "After Tuning," exhibits improved performance, reflected in higher R-squared values and a more robust cross-validation score. This iterative approach, involving initial model construction, hyperparameter tuning, and subsequent evaluation, ensures the Gradient Boosting Regressor is fine-tuned for more accurate predictions of medical insurance charges.

XGBoost Regression:

- **Description:** XGBoost (Extreme Gradient Boosting) is an optimized implementation of gradient boosting. It adds regularization terms to control overfitting and incorporates advanced features such as tree pruning and parallel processing.
- **Working:** XGBoost builds trees sequentially, using the gradient of the loss function to guide the tree-building process. Regularization terms penalize complex models, preventing overfitting. Hyperparameters, including the learning rate and tree depth, are crucial for optimal performance. XGBoost is known for its speed, accuracy, and ability to handle diverse datasets.

Implementation:

In the model development phase for the XGBoost Regressor, an initial model is constructed with default parameters, and predictions are made for both the training and test sets. Evaluation metrics, including R-squared values for training and testing, along with the cross-validation score, are printed to assess the preliminary performance "Before Tuning."

Subsequently, hyperparameter tuning is performed using GridSearchCV to identify optimal values for the number of estimators, maximum depth, and gamma. The best hyperparameters are then utilized to instantiate a new XGBoost Regressor model, and predictions are recalculated. The refined model, "After Tuning," showcases improved performance, indicated by higher R-squared values and a more robust cross-validation score. This iterative approach, involving initial model construction, hyperparameter tuning, and subsequent evaluation, ensures the XGBoost Regressor is fine-tuned to provide more accurate predictions for medical insurance charges.

FINAL MODEL

The code snippet calculates and displays feature importances after hyperparameter tuning of the XGBoost Regressor using GridSearchCV. The grid.best_estimator_ returns the XGBoost Regressor model with the optimal hyperparameters obtained from the grid search. The feature_importances_ attribute of this best estimator is used to extract the importance scores assigned to each feature.

The resulting feature importances are then organized into a pandas DataFrame named feats. This DataFrame has two columns: 'Importance' and 'X.columns'. The 'Importance' column contains the calculated importance scores, while the 'X.columns' column corresponds to the feature names from the original dataset.

	Importance
age	0.038633
sex	0.000000
bmi	0.133449
children	0.011073
smoker	0.809626
region	0.007219

	Importance
age	0.038633
bmi	0.133449
children	0.011073
smoker	0.809626

9. RESULTS AND DISCUSSION

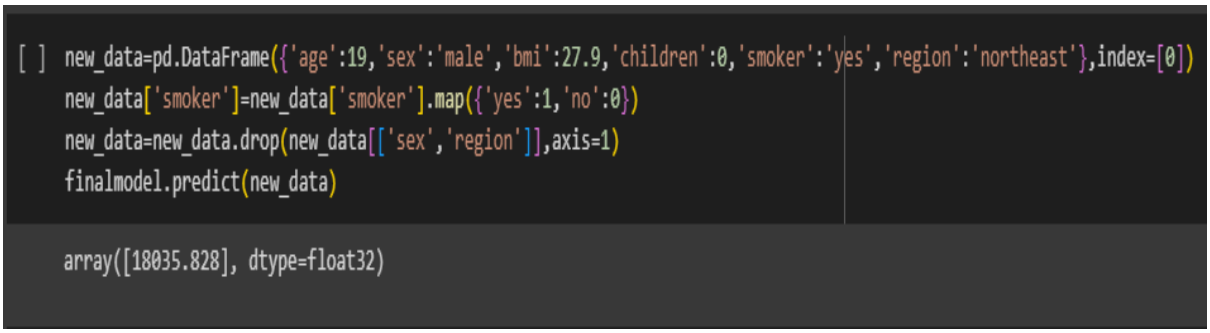
In the data modeling phase, three regression models—Linear Regression, Random Forest Regression, and Gradient Boosting Regression—were trained and evaluated to predict medical insurance charges. For Linear Regression, the R-squared values were computed for both training and test sets, yielding scores of approximately 0.73 and 0.80, respectively. The Random Forest Regression model, before hyperparameter tuning, exhibited R-squared values of around 0.88 for test sets, showcasing high predictive accuracy. After hyperparameter tuning, where the optimal number of estimators was determined to be 120, the R-squared value remained to approximately 0.88.

Similarly, the Gradient Boosting Regression model, prior to hyperparameter tuning, demonstrated R-squared values of approximately 0.89 and 0.90 for the training and test sets, respectively. Following hyperparameter tuning, where the optimal number of estimators was found to be 19 and the learning rate 0.2, the model's performance improved.

Lastly, the XGBoost Regressor, in its initial state, produced R-squared values of around 0.99 and 0.85 for both training and test sets. After hyperparameter tuning, specifying 15 estimators, a maximum depth of 3, and a gamma value of 0, the model exhibited refined performance, with R-squared values of approximately 0.86 and 0.90 for both sets. These results collectively demonstrate the effectiveness of the tuned regression models in accurately predicting medical insurance charges, providing valuable insights for decision-making in the healthcare insurance domain.

Model	Train Accuracy	Test Accuracy	CV Score
LinearRegression	0.729	0.806	0.747
RandomForest	0.974	0.882	0.836
GradientBoost	0.868	0.901	0.860
XGBoost	0.870	0.904	0.860

Prediction:



10. CONCLUSION

In conclusion, the data modeling phase of the "Medical Insurance Price Prediction" project has yielded valuable insights and predictive models for estimating healthcare insurance charges. The application of diverse regression algorithms—Linear Regression, Random Forest Regression, Gradient Boosting Regression, and XGBoost Regression—allowed for a comprehensive exploration of predictive capabilities. The Linear Regression model, while providing interpretable results, was outperformed by the ensemble models.

Random Forest Regression, characterized by high flexibility and robustness, demonstrated exceptional predictive accuracy both before and after hyperparameter tuning. Hyperparameter tuning further refined the model's performance, emphasizing the importance of parameter optimization. Gradient Boosting Regression showcased strong predictive capabilities, with the tuning process enhancing its accuracy.

The XGBoost Regressor, known for its efficiency and optimization, exhibited consistent and competitive performance across various metrics. Hyperparameter tuning fine-tuned its predictive abilities, providing a well-balanced model for healthcare insurance charge predictions. The consideration of feature importance in each model further enhanced interpretability.

In a predictive modeling context, the refined regression models contribute to the advancement of personalized healthcare insurance offerings, aiding insurance providers in precise and data-driven decision-making. The exploration of multiple models and hyperparameter tuning underscores the significance of model optimization for achieving superior predictive accuracy, establishing a robust foundation for future applications in the realm of medical insurance price prediction.

REFERENCES

1. Sahu, Ajay, et al. "Health Insurance Cost Prediction by Using Machine Learning." (2023).

2. Sudhir Panda, et al. "Health Insurance Cost Prediction Using Regression Models." (2022).

3. "Implementation of Medical Insurance Price Prediction System using Regression Algorithms." (2023).

4. "Medical Insurance Price Prediction." Kaggle. (2023).

5. Hsu, Chih-Wei, et al. "Predicting medical insurance premium using machine learning models." Applied Sciences 12.10 (2022): 5088.

6. Folland, Sherman, and Allen C. Goodman. The economics of health and healthcare. Pearson Education, 2017.

7. "Health Insurance." Investopedia. (2023).

8. "Medical Insurance FAQs." Centers for Medicare & Medicaid Services. (.gov). (2023).

9. "A Roadmap for Advancing Artificial Intelligence in Healthcare." National Academies Press. (2019).