

Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Insurance Price prediction



Supervised By:

Mrs. Htet Ne Oo

Submitted By:

Rajkirat Singh 2210990709

Rajveer Singh 2210990711

Ranveer Singh 2210990718

Rohit Chauhan 2210990742

**Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab**

Abstract

This project aims to develop a predictive model for insurance pricing using machine learning techniques. By leveraging historical data and relevant features, the model can accurately estimate insurance premiums, enabling insurance companies to optimize their pricing strategies and provide fair quotes to customers. The project explores various algorithms, feature engineering techniques, and model evaluation metrics to achieve a robust and reliable solution.

Table Of Contents

| S.No | Topics | Page No. |
|-------|-------------------------------------|----------|
| 1. | Introduction | 4-8 |
| ➤ 1.1 | ➤ Background | |
| ➤ 1.2 | ➤ Objectives | |
| ➤ 1.3 | ➤ Significance | |
| 2. | Problem Definition and Requirements | 9 |
| ➤ 2.1 | ➤ Problem Statement | |
| ➤ 2.2 | ➤ Data Requirements | |
| ➤ 2.3 | ➤ Software Requirements | |
| 3. | Proposed Design/Methodology | 10-11 |
| ➤ 3.1 | ➤ Data Preprocessing | |
| ➤ 3.2 | ➤ Feature Engineering | |
| ➤ 3.3 | ➤ Model Selection and Training | |
| ➤ 3.4 | ➤ Model Evaluation and Optimization | |
| ➤ 3.5 | ➤ Deployment and Integration | |
| 4. | Results | 12-15 |

1. Introduction

1.1 Background

Insurance pricing is a critical function within the insurance industry, as it directly impacts the profitability and competitiveness of insurance companies while also determining the affordability and accessibility of insurance products for consumers. Traditionally, insurance pricing has been governed by actuarial models and statistical techniques that rely heavily on historical claims data and predefined risk factors.

Actuarial models involve the analysis of past claims data, risk factors, and probability distributions to estimate the likelihood and potential cost of future claims. These models consider various factors such as age, gender, location, driving record, and vehicle information to calculate insurance premiums. However, these traditional methods often struggle to capture the complex relationships and non-linear patterns present in the data, leading to potential inaccuracies or biases in pricing.

With the advent of machine learning and artificial intelligence (AI) techniques, there is a growing opportunity to develop more advanced and accurate insurance pricing models. Machine learning algorithms have the ability to learn from large datasets, identify intricate patterns, and make data-driven predictions with higher accuracy compared to traditional statistical methods.

Some of the key advantages of using machine learning for insurance pricing include:

1. **Handling High-Dimensional Data:** Machine learning models can effectively handle and process large volumes of data with numerous features, capturing complex interactions and non-linear relationships that may be difficult to identify through traditional methods.
2. **Continuous Learning and Adaptation:** Machine learning models can be retrained and updated as new data becomes available, enabling continuous improvement and adaptation to changing market conditions and risk factors.
3. **Personalized Pricing:** By leveraging a wide range of features and individual characteristics, machine learning models can provide more personalized and accurate

pricing for each policyholder, ensuring fairness and reducing the potential for over-pricing or under-pricing.

4. **Predictive Power:** Advanced machine learning algorithms, such as neural networks and ensemble methods, can uncover intricate patterns and make highly accurate predictions, leading to improved risk assessment and more competitive pricing strategies.
5. **Automation and Scalability:** Machine learning models can be integrated into automated pricing systems, enabling insurance companies to streamline their pricing processes, reduce manual effort, and scale their operations more efficiently.

1.2 Objectives

The primary objective of this project is to develop a robust and accurate machine learning model for predicting insurance prices. This overarching goal can be further divided into the following specific objectives:

1. **Data Preprocessing and Feature Engineering:** One of the key objectives is to preprocess the available insurance data effectively. This involves handling missing values, encoding categorical variables, and scaling numerical features to ensure the data is in a suitable format for machine learning algorithms. Additionally, the project aims to explore various feature engineering techniques, such as creating new features from existing ones or selecting the most relevant features, to enhance the model's predictive performance.
2. **Algorithm Selection and Model Training:** Another objective is to evaluate and compare the performance of various machine learning algorithms for insurance price prediction. This includes algorithms such as linear regression, decision trees, random forests, gradient boosting, and neural networks. The project aims to train these algorithms on the preprocessed dataset and assess their performance using appropriate evaluation metrics, such as mean squared error, mean absolute error, and R-squared.
3. **Model Optimization and Hyperparameter Tuning:** To achieve the best possible predictive performance, the project aims to optimize the selected machine learning model(s) through hyperparameter tuning. This involves systematically exploring different

combinations of hyperparameters to find the optimal configuration that maximizes the model's accuracy and generalization capability.

4. **Cross-Validation and Generalization:** Ensuring the model's robustness and ability to generalize to unseen data is crucial. The project objective includes employing cross-validation techniques, such as k-fold cross-validation or stratified cross-validation, to assess the model's performance on different subsets of the data and mitigate overfitting.
5. **Interpretability and Explainability:** While achieving high predictive accuracy is essential, the project also aims to explore methods for interpreting and explaining the machine learning model's predictions. This can involve techniques like feature importance analysis, partial dependence plots, or SHAP (SHapley Additive exPlanations) values, which can provide insights into the model's decision-making process and the relative importance of different features.
6. **Integration and Deployment:** The ultimate objective is to deploy and integrate the developed machine learning model into the insurance company's existing systems. This may involve creating user interfaces, APIs, or batch processing pipelines to facilitate the usage of the model for insurance pricing in real-world scenarios.
7. **Ethical Considerations:** Throughout the project, ethical considerations will be taken into account to ensure fairness, transparency, and accountability in the insurance pricing process. The project aims to identify and mitigate potential biases or discriminatory practices that may arise from the use of machine learning models.

1.3 Significance

The development of an accurate and robust machine learning model for insurance price prediction holds significant implications for both insurance companies and policyholders. The successful implementation of such a model can have far-reaching impacts on various aspects of the insurance industry, including profitability, risk management, customer satisfaction, and market competitiveness.

1. **Improved Profitability for Insurance Companies:** Accurate insurance pricing is crucial for insurance companies to maintain profitability and long-term sustainability. By leveraging machine learning models, insurance companies can better assess risk factors,

identify premium pricing opportunities, and optimize their pricing strategies. This can lead to increased revenue, reduced losses, and improved overall financial performance.

2. **Enhanced Risk Management:** Insurance companies rely on effective risk assessment and management to ensure their solvency and ability to meet claims obligations. Machine learning models can provide more accurate risk predictions by analyzing complex data patterns and capturing non-linear relationships that may be difficult to identify through traditional actuarial methods. This improved risk assessment can help insurance companies better manage their risk exposure and make more informed underwriting decisions.
3. **Fair and Transparent Pricing for Policyholders:** Traditional insurance pricing methods may sometimes result in over-pricing or under-pricing for certain customer segments due to inherent biases or limitations in the actuarial models. Machine learning models, when developed and implemented responsibly, can provide more personalized and fair pricing by considering a wide range of relevant factors and individual characteristics. This can lead to increased transparency and trust between insurance companies and policyholders.
4. **Competitive Advantage in the Market:** The ability to accurately price insurance products is a key competitive advantage in the highly competitive insurance market. Companies that leverage machine learning for insurance pricing can gain a significant edge over their competitors by offering more attractive and tailored pricing options to customers, potentially increasing their market share and customer loyalty.
5. **Operational Efficiency and Scalability:** Machine learning models can be integrated into automated pricing systems, enabling insurance companies to streamline their pricing processes, reduce manual effort, and scale their operations more efficiently. This can lead to cost savings, faster turnaround times, and improved customer experiences.
6. **Data-Driven Decision Making:** The implementation of machine learning models in insurance pricing promotes a data-driven approach to decision-making within the organization. By relying on empirical data and advanced analytical techniques, insurance companies can make more informed and objective decisions, reducing the reliance on subjective judgments or outdated practices.
7. **Regulatory Compliance and Ethical Considerations:** As the use of machine learning in insurance pricing gains traction, regulatory bodies and ethical guidelines are likely to evolve to ensure fairness, transparency, and accountability. By proactively addressing

these concerns and adhering to best practices, insurance companies can stay ahead of regulatory changes and maintain a positive public perception.

2. Problem Definition and Requirements

2.1 Problem Statement

The problem addressed in this project is the development of a machine learning model that can accurately predict insurance prices based on various features and historical data. The model should be able to learn from past data and make reliable predictions for new customers or policies.

2.2 Data Requirements

To train and evaluate the machine learning model, a comprehensive dataset containing historical insurance data is required. The dataset should include relevant features such as age, gender, vehicle information, driving history, and previous claims, as well as the corresponding insurance prices.

2.3 Software Requirements

The project will utilize Python programming language and various machine learning libraries such as scikit-learn, TensorFlow, or PyTorch. Additionally, data preprocessing, visualization, and evaluation libraries like Pandas, NumPy, and Matplotlib may be employed.

3. Proposed Methodology

3.1 Data Preprocessing

The first step in the methodology involves data preprocessing, which includes handling missing values, encoding categorical variables, and scaling numerical features. This ensures that the data is in a suitable format for machine learning algorithms.

3.2 Feature Engineering

Feature engineering techniques, such as creating new features from existing ones or selecting the most relevant features, will be explored to improve the model's predictive performance.

3.3 Model Selection and Training

Various machine learning algorithms will be evaluated, including linear regression, decision trees, random forests, gradient boosting, and neural networks. These algorithms will be trained on the preprocessed dataset, and their performance will be assessed using appropriate evaluation metrics.

3.4 Model Evaluation and Optimization

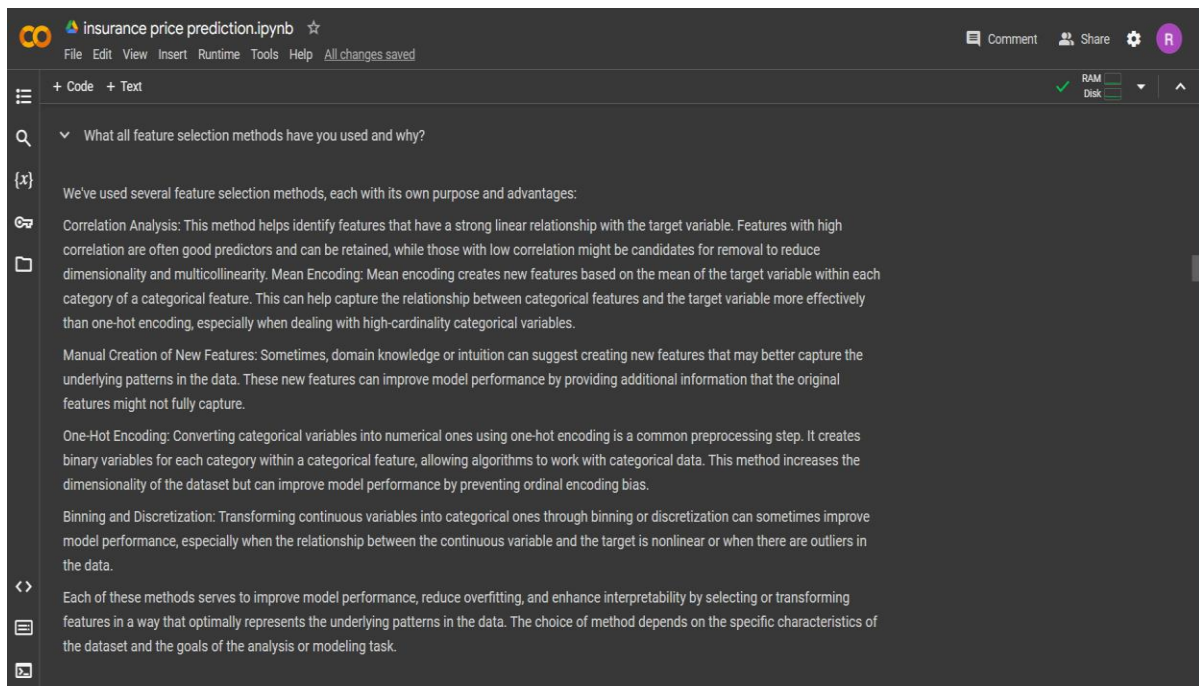
The trained models will be evaluated using metrics such as mean squared error, mean absolute error, and R-squared. Cross-validation techniques will be employed to ensure the model's robustness and generalization capability. Hyperparameter tuning and ensemble methods may be explored to optimize the model's performance further.

3.5 Deployment and Integration

Once the best-performing model is selected, it will be deployed and integrated into the insurance company's existing systems. This may involve creating user interfaces, APIs, or batch processing pipelines to facilitate the usage of the model for insurance pricing.

4. Results

This section will present the findings of the project, including the performance metrics of the final model, visualizations, and analyses. It will also discuss the model's limitations and potential areas for future improvement.



The screenshot shows a Jupyter Notebook interface with the title "insurance price prediction.ipynb". The notebook is in "Text" mode. The content of the cell is as follows:

What all feature selection methods have you used and why?

We've used several feature selection methods, each with its own purpose and advantages:

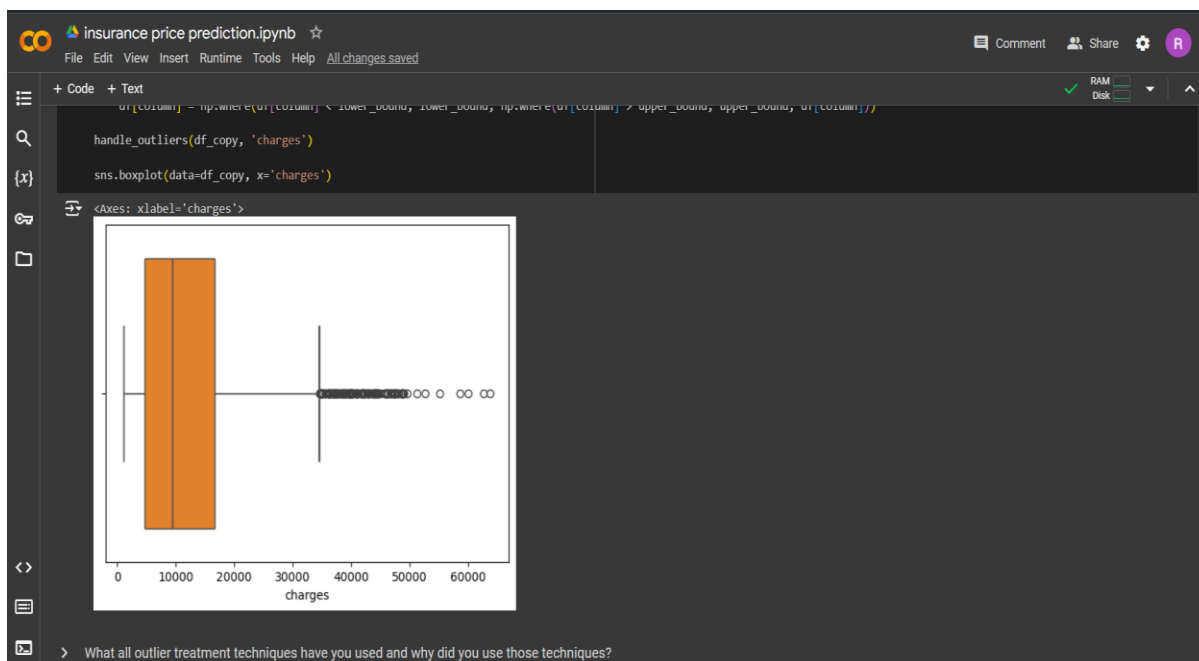
Correlation Analysis: This method helps identify features that have a strong linear relationship with the target variable. Features with high correlation are often good predictors and can be retained, while those with low correlation might be candidates for removal to reduce dimensionality and multicollinearity. Mean Encoding: Mean encoding creates new features based on the mean of the target variable within each category of a categorical feature. This can help capture the relationship between categorical features and the target variable more effectively than one-hot encoding, especially when dealing with high-cardinality categorical variables.

Manual Creation of New Features: Sometimes, domain knowledge or intuition can suggest creating new features that may better capture the underlying patterns in the data. These new features can improve model performance by providing additional information that the original features might not fully capture.

One-Hot Encoding: Converting categorical variables into numerical ones using one-hot encoding is a common preprocessing step. It creates binary variables for each category within a categorical feature, allowing algorithms to work with categorical data. This method increases the dimensionality of the dataset but can improve model performance by preventing ordinal encoding bias.

Binning and Discretization: Transforming continuous variables into categorical ones through binning or discretization can sometimes improve model performance, especially when the relationship between the continuous variable and the target is nonlinear or when there are outliers in the data.

Each of these methods serves to improve model performance, reduce overfitting, and enhance interpretability by selecting or transforming features in a way that optimally represents the underlying patterns in the data. The choice of method depends on the specific characteristics of the dataset and the goals of the analysis or modeling task.

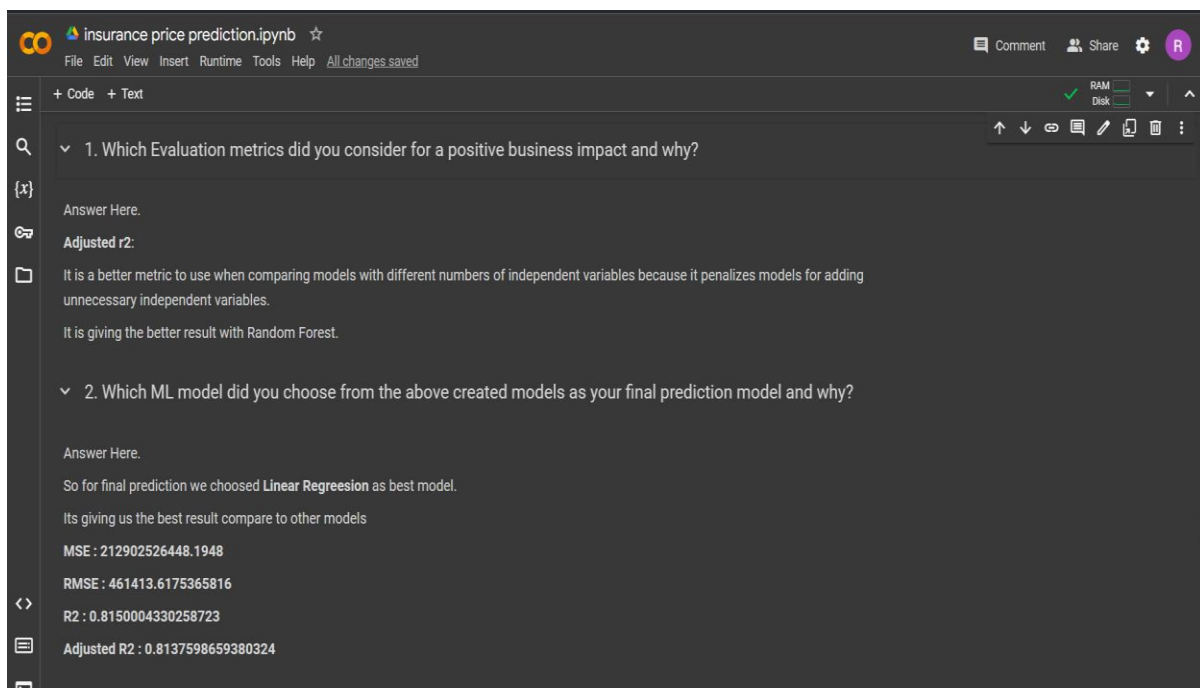
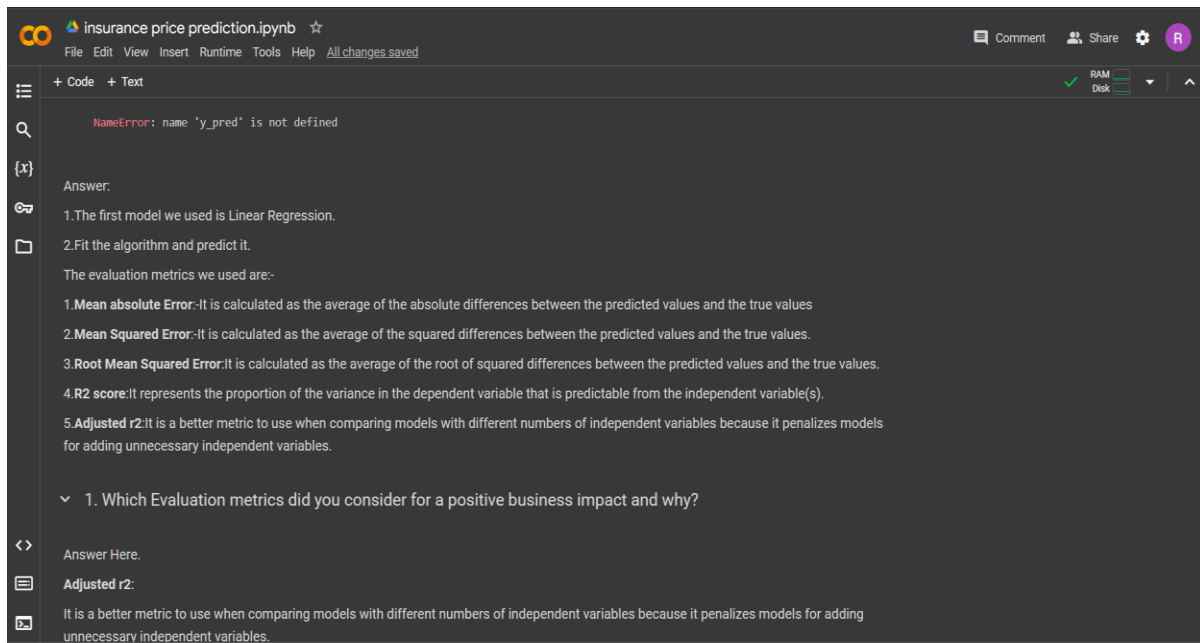


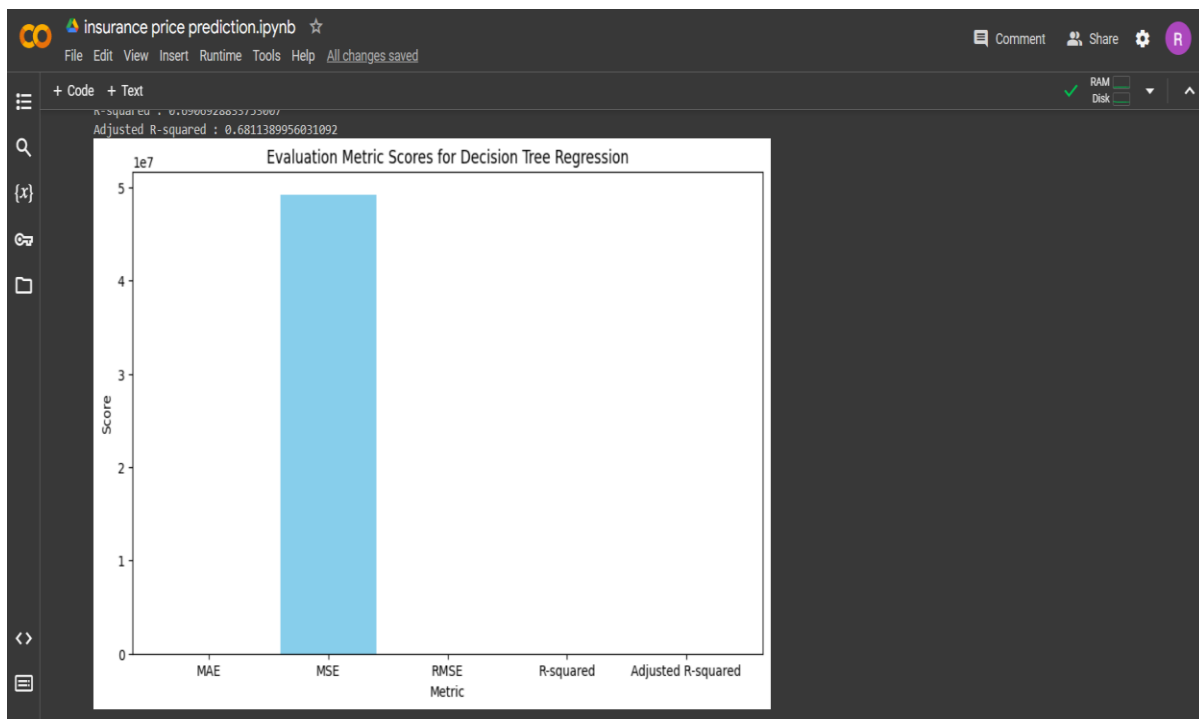
The screenshot shows a Jupyter Notebook interface with the title "insurance price prediction.ipynb". The notebook is in "Code" mode. The code cell contains the following code:

```
df[column] = np.where(df[column] < lower_bound, lower_bound, np.where(df[column] > upper_bound, upper_bound, df[column]))  
handle_outliers(df_copy, 'charges')  
sns.boxplot(data=df_copy, x='charges')
```

The output of the code is a boxplot titled "charges". The x-axis is labeled "charges" and ranges from 0 to 60,000. The y-axis is labeled "charges". The boxplot shows a distribution of charges with a median around 10,000, a mean around 15,000, and several outliers extending up to 60,000.

What all outlier treatment techniques have you used and why did you use those techniques?





When evaluating the performance of a machine learning model for insurance price prediction, several metrics can be used to assess its accuracy and effectiveness. Here are some commonly used metrics in this context:

1. **Mean Squared Error (MSE):** This metric measures the average squared difference between the predicted insurance prices and the actual prices. It is calculated as the sum of squared differences divided by the number of observations. A lower MSE value indicates better model performance. $MSE = (1/n) \sum (y_i - \hat{y}_i)^2$ Where:
 - n is the number of observations
 - y_i is the actual insurance price for observation i
 - \hat{y}_i is the predicted insurance price for observation i
2. **Root Mean Squared Error (RMSE):** The RMSE is the square root of the MSE and provides a measure of the average magnitude of the errors in the same unit as the target variable (insurance price). Lower RMSE values indicate better model performance. $RMSE = \sqrt{MSE}$

3. **Mean Absolute Error (MAE):** The MAE measures the average absolute difference between the predicted insurance prices and the actual prices. It is less sensitive to outliers compared to MSE and RMSE. $MAE = (1/n) \sum |y_i - \hat{y}_i|$
4. **R-squared (R^2):** R-squared is a statistical measure that represents the proportion of the variance in the dependent variable (insurance price) that is explained by the independent variables (features) in the model. It ranges from 0 to 1, with higher values indicating better model fit. $R^2 = 1 - (\sum (y_i - \hat{y}_i)^2) / (\sum (y_i - \bar{y})^2)$ Where \bar{y} is the mean of the actual insurance prices.
5. **Mean Absolute Percentage Error (MAPE):** MAPE is a relative metric that measures the average absolute difference between the predicted and actual insurance prices as a percentage of the actual prices. It is useful when the scale of the target variable is important. $MAPE = (1/n) \sum (|y_i - \hat{y}_i| / y_i) * 100\%$
6. **Residual Plots:** Residual plots can be used to visualize the difference between the predicted and actual insurance prices. These plots can help identify patterns or systematic errors in the model's predictions.
7. **Feature Importance:** Techniques like permutation importance or SHAP (SHapley Additive exPlanations) values can be used to understand the relative importance of different features in the model's predictions. This information can provide insights into the most influential factors for insurance pricing.