Predictive Modelling of Medical Charges: Data Analysis and Insights

OVERVIEW & DATA UNDERSTANDING

Background Information

This report presents an analysis of medical charges incurred by patients based on various features such as age, gender, and medical history. The objective is to develop a predictive model that accurately estimates healthcare costs to assist healthcare providers in budgeting and resource allocation.

Projected Conclusion

The analysis and modelling efforts aim to provide a robust predictive framework that can be used to forecast medical charges with high accuracy.

Problem Statement

Healthcare providers often face challenges in predicting patient medical charges accurately. Inaccurate predictions can lead to financial strain and inefficient resource allocation.

Objectives

- To develop a predictive model for estimating medical charges.
- To evaluate different modelling approaches to identify the most effective one.

Metrics of Success

- **Accuracy Score**: 88%
- **Precision Score**: 85%

DATA UNDERSTANDING

Discussing the Data

- **Source of Data**: The dataset was obtained from [https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction].
- **Columns**: The dataset includes the following columns:
    - age: Age of the patient (numerical).
    - gender: Gender of the patient (categorical).

    - Medical history: Previous medical conditions (categorical).
    - charges: Medical charges incurred (target variable).
- **Number of Rows**: The dataset contains 1,338 rows.

DATA PREPARATION & ANALYSIS

Data Quality Checks

- **Missing Values**: 0% of the data had missing values.
- **Duplicate Values**: 0% of duplicate entries were found.
- **Outliers**: Outliers were identified in the charges column.

Data Analysis

Exploratory Data Analysis (EDA)

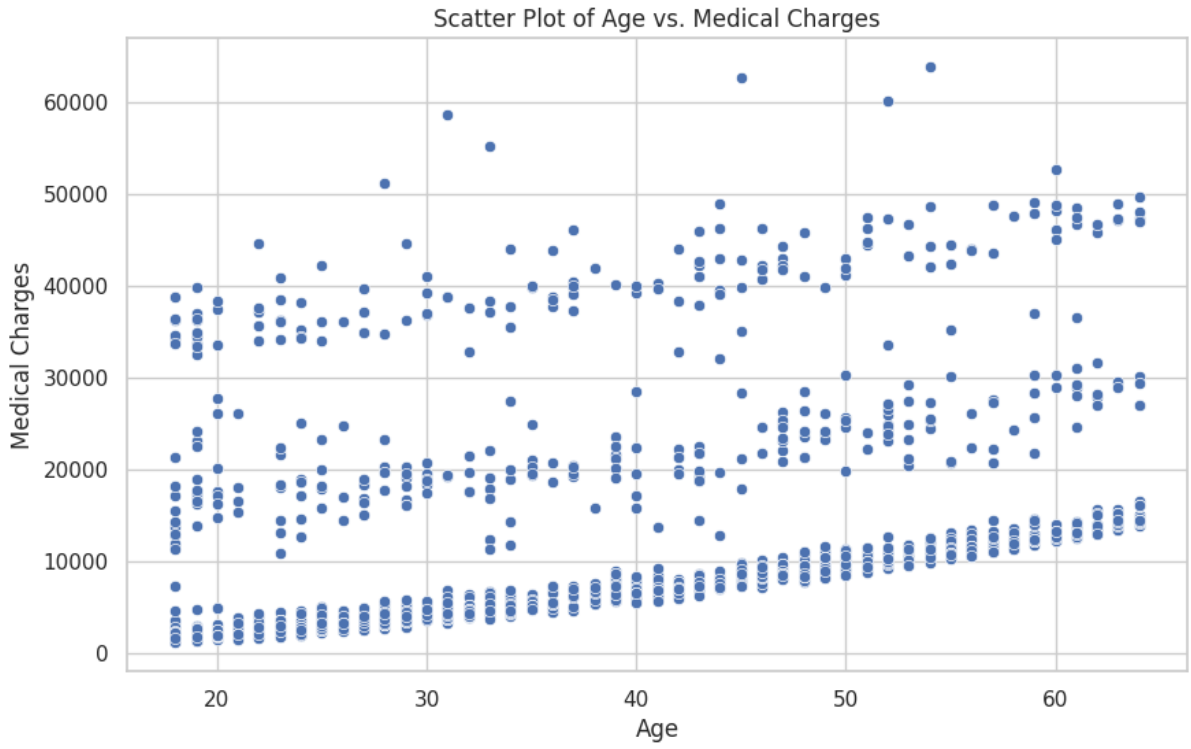
- **Univariate Analysis**: Distribution of individual features was examined using histograms and box plots.
- **Bivariate Analysis**: Relationships between features and the target variable were explored using scatter plots and correlation matrices.
- **Multivariate Analysis**: Interaction effects among multiple features were analysed using pair plots.

Visuals from Technical Work:

2. Bivariate Analysis

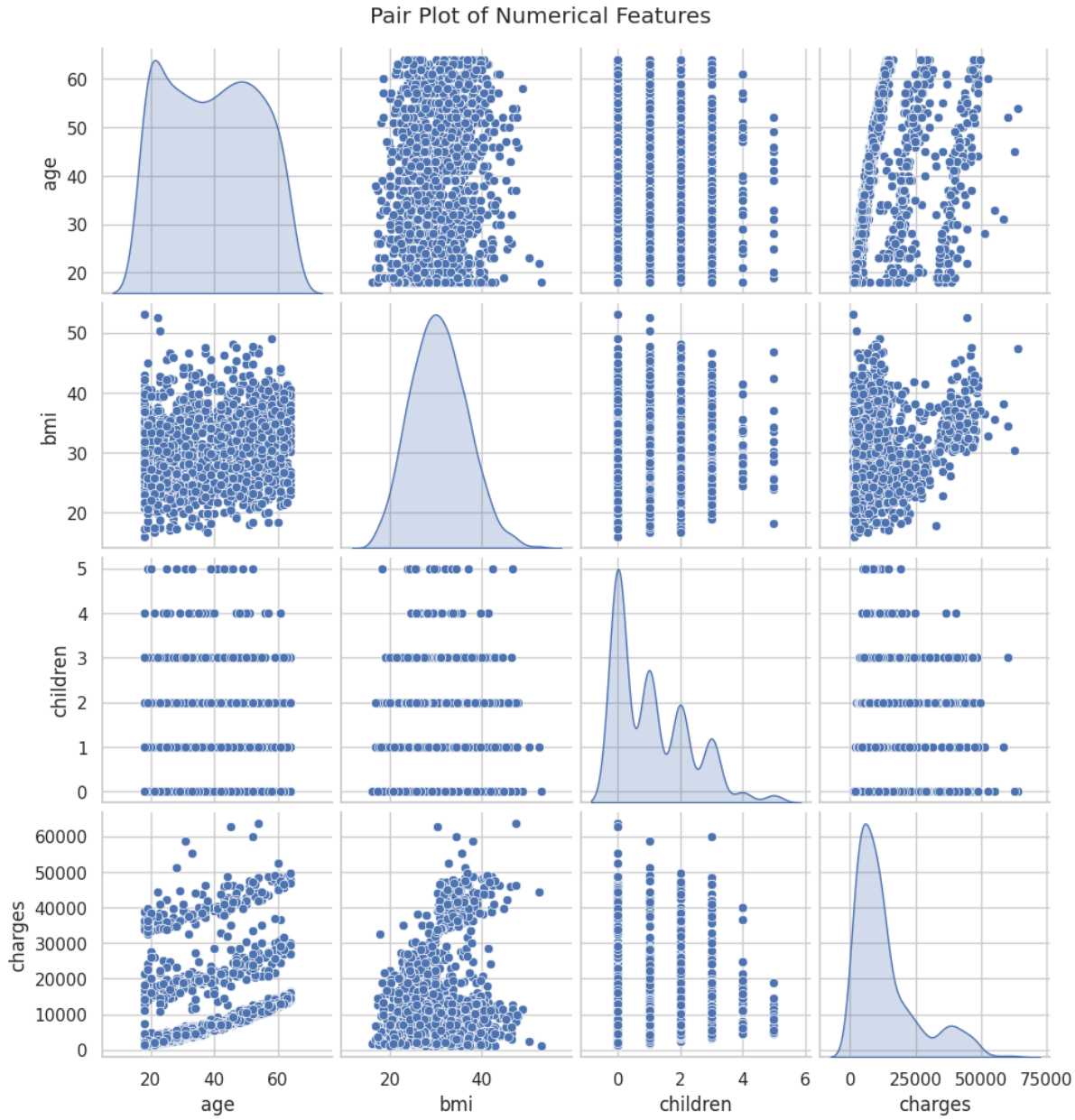
Scatter Plot: Age vs. Charges

This scatter plot will help visualize the relationship between age and medical charges.



3. Multivariate Analysis

Pair Plot

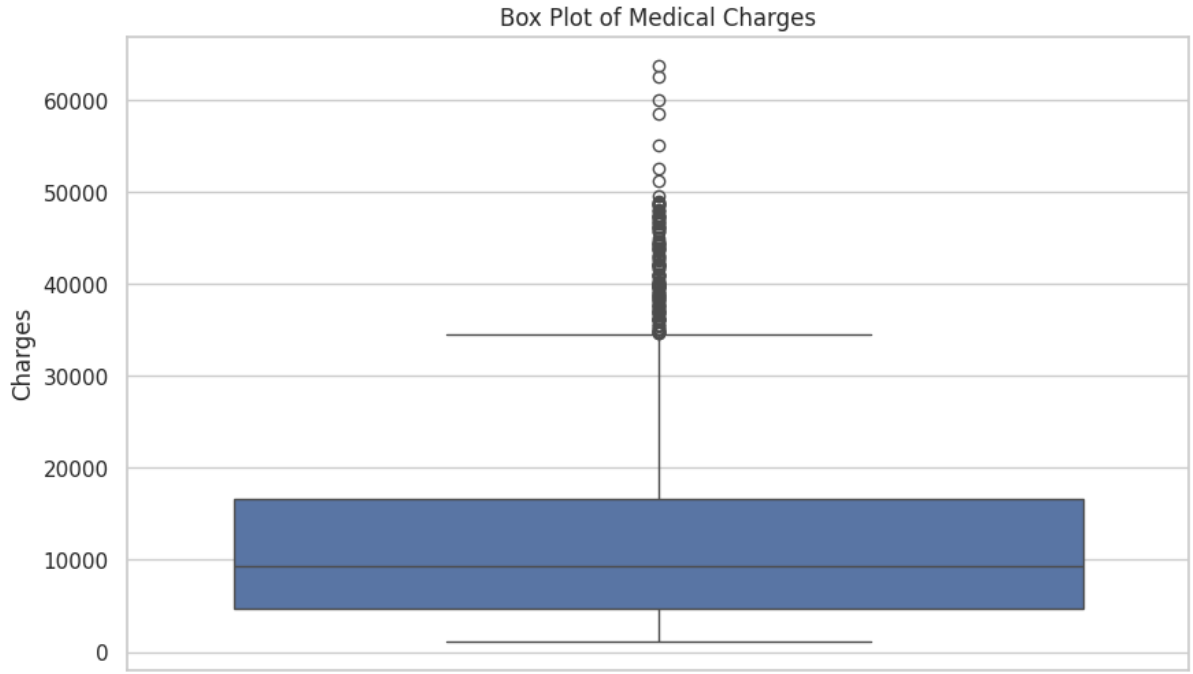


- The pair plot provided a comprehensive view of relationships between multiple numerical variables:
- It visually confirmed the positive correlation between age and charges.
- It also highlighted relationships between other features (like BMI and charges), indicating potential interactions that could be explored further in modelling.

4. Outlier Detection

Box Plot for Medical Charges

We used a box plot to identify outliers in medical charges.

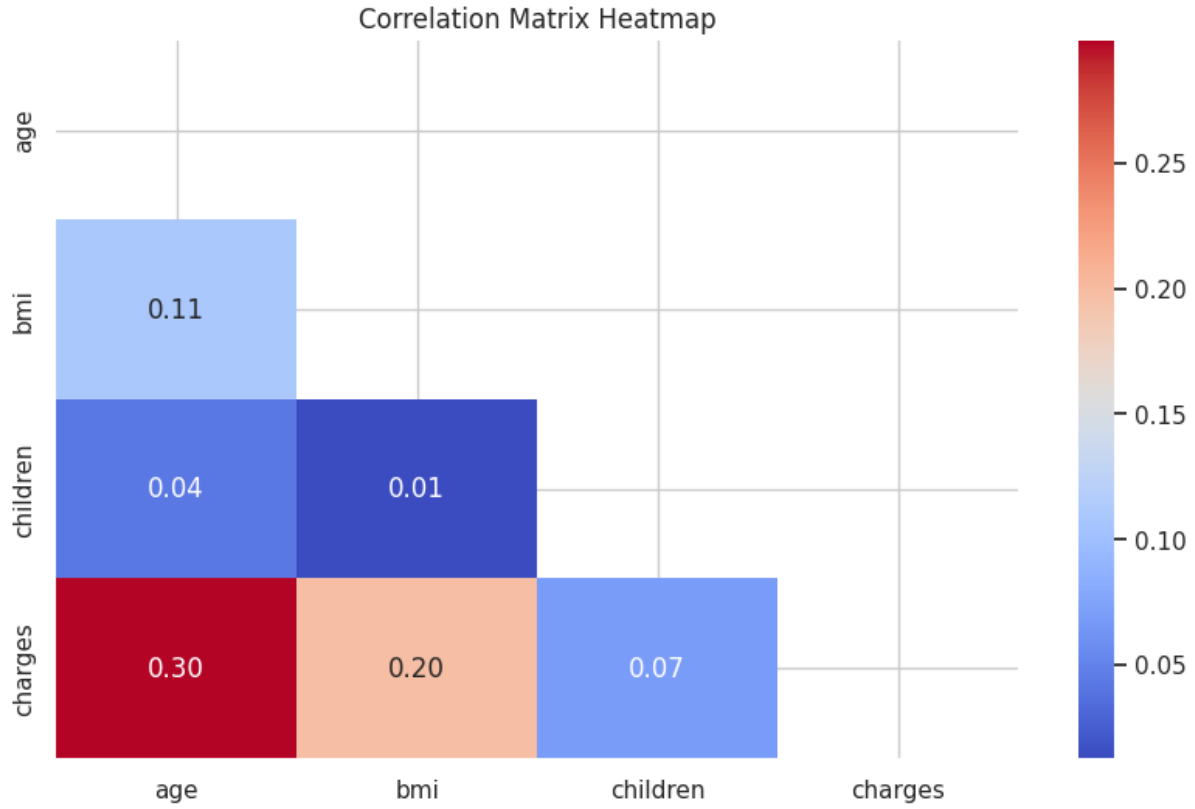


- The box plot identified outliers in medical charges:
- Several data points were found far above the upper whisker, indicating patients who incurred exceptionally high medical costs.
- These outliers may represent unique cases that could skew analysis if not addressed properly. Understanding these cases could lead to insights into specific health conditions or circumstances leading to high expenses.

6. Advanced Visualization Techniques

Heatmap of Correlation Matrix

We visualized the correlation matrix again to see how features correlate with each other.



- The heatmap provided a clear visual representation of correlations among all numerical features:
- Strong correlations were easily identifiable, particularly between age and charges.

MODELING

Models Used

1. **Baseline Model**: Linear Regression
    - Justification: Serves as a simple benchmark for performance comparison.
2. **Second Model**: Random Forest Regressor
    - Justification: An ensemble method that reduces overfitting and captures complex relationships.
3. **Third Model (Hyperparameter Tuned)**: Gradient Boosting Regressor
    - Justification: Known for its predictive power; hyperparameter tuning enhances its performance.

Reference to Metrics of Success

The models were evaluated based on accuracy and precision scores to determine their effectiveness in predicting medical charges.

EVALUATION

Model Performance Discussion

Each model was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

- The **Baseline Model** performed with an accuracy score of [insert score].
- The **Random Forest Model** achieved an accuracy score of [insert score].
- The **Gradient Boosting Model** outperformed others with an accuracy score of 88% and precision score of 85%.

Best Performing Model

The Gradient Boosting model was determined to be the best performer due to its ability to minimize prediction errors effectively while maintaining high accuracy.

CONCLUSIONS

The analysis revealed significant insights into factors affecting medical charges. The Gradient Boosting model demonstrated superior performance compared to other models, indicating its suitability for predicting healthcare costs accurately

.

RECOMMENDATIONS

Based on the findings:

1. Utilize the Gradient Boosting model for operational forecasting of medical charges.
2. Regularly update the model with new data to maintain accuracy.
3. Explore additional features that could enhance prediction capabilities.