Report - Group 13

Group:

- Group Number 13

Group Members:

- Nanditha Chevula
- Netra Amin
- Yusuf Kshem
- Anudeepthi Senthil Kumar

Dataset:

We used the Medical Insurance Cost Prediction dataset:
https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction

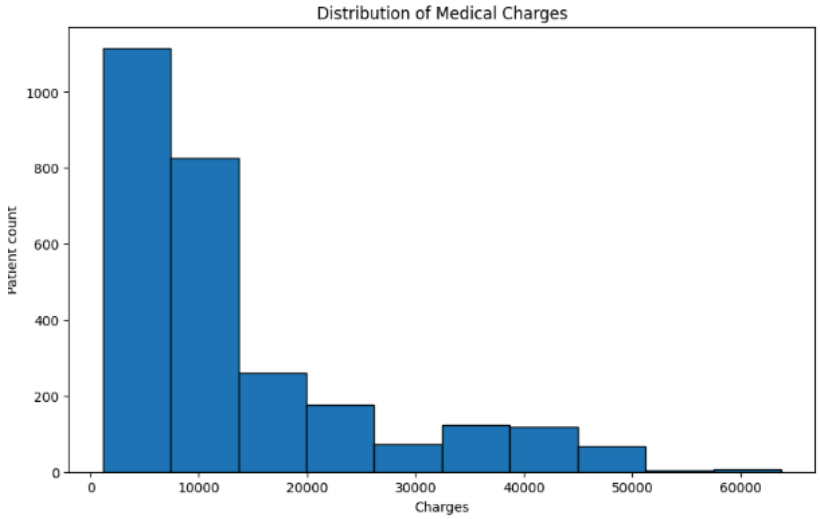
This dataset contains demographic and lifestyle factors that influence medical expenses, including age, sex, BMI, smoking status, number of children, and region. Using these features, we are attempting to predict insurance price, which represents individual medical insurance costs. There are 2,700 rows and 7 columns which are a combination of categorical and numerical data.

Our goal is to predict medical insurance charges based on the input features. Each group member implemented and analyzed one model.

Exploratory Data Analysis:

- For summary statistics, we used the following:
    - Mean of charges and bmi
        - Mean of charges: $13,261.37
        - Mean of BMI: 30.70
    - Median of charges and bmi
        - Median of charges: $9,333.01
        - Median of BMI: 30.45
        - Since the mean of charges is higher than the median of charges, it suggests that there might be some outliers in the dataset and that the data overall might be right skewed. The mean and median of BMI seems relatively close so it might be normally distributed.
    - Standard deviation of charges and bmi
        - Standard Deviation of charges: $12,151.77
        - Standard Deviation of BMI: 6.13

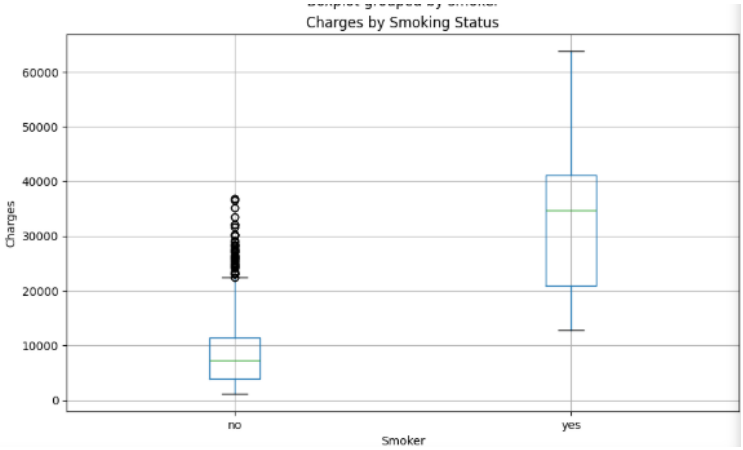
        - This is a very large standard deviation for charges, indicating that medical charges vary widely. BMI on the other hand seems moderately small.
    - Count of categorical variables such as sex, smoker, and region
        - Count for sex:
            - Male = 1406
            - Female = 1,366
        - Count for smoker:
            - non-smoker= 2,208
            - Smoker = 564
        - Count for region:
            - southeast = 766
            - southwest = 684
            - northwest = 664
            - northeast = 658
        - The Sex class seems nearly distributed with no class imbalances since it is almost split 50-50. The smoker class on the other hand is uneven. There are far more non-smokers which might affect how the model learns smoking related influence on the charges. Region class seems very fairly balanced.
    - Minimum and maximum values of charges
        - Min of charges = $1,121.87
        - Max of charges = $63,770.43
        - Seeing that it is a very wide range, we can assume that there might be presence of some outliers in medical charges in the dataset.
- Visualizations
    - Histogram of charges



        - Helps to actually visualize the distribution of medical charges. and detect skewness or outliers. Based on the graph, it tells us that medical insurance

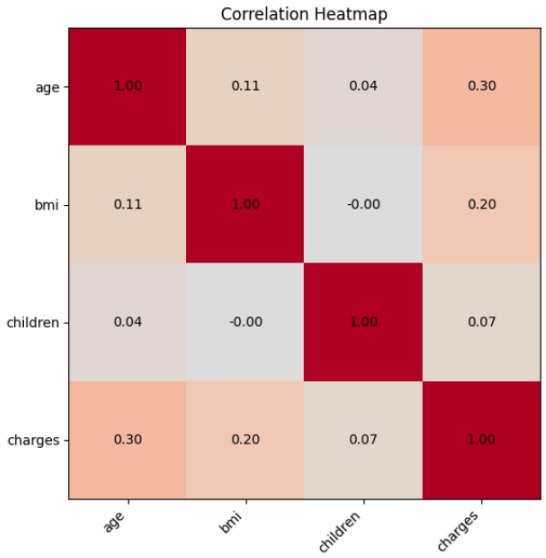
charges are heavily skewed to the right. Very small numbers of people tend to have high charges over $30000. We can also assume there might be presences of outliers based on the overall distribution of the dataset.

- Boxplot of charges by smoking status



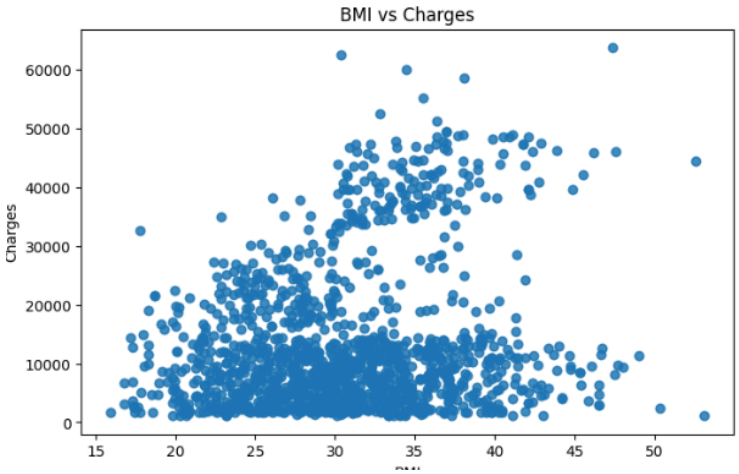
  - To compare cost distributions between smokers and non-smokers. This plot provides us with evidence that smokers then have higher charges and smoking is a highly influential factor in predicting medical charges. Individuals who smoke then have higher charges, which can be seen by looking at the medians for both non smokers and smokers. The median for charges for smokers is well above the highest quartile for non-smokers. Smokers have a larger IQR, meaning they have more variability in charges. On the other hand, non-smokers have a tighter IQR. Overall, the plot tells us that smoking seems to be a major driver of increased medical costs.

- Heatmap of feature correlations



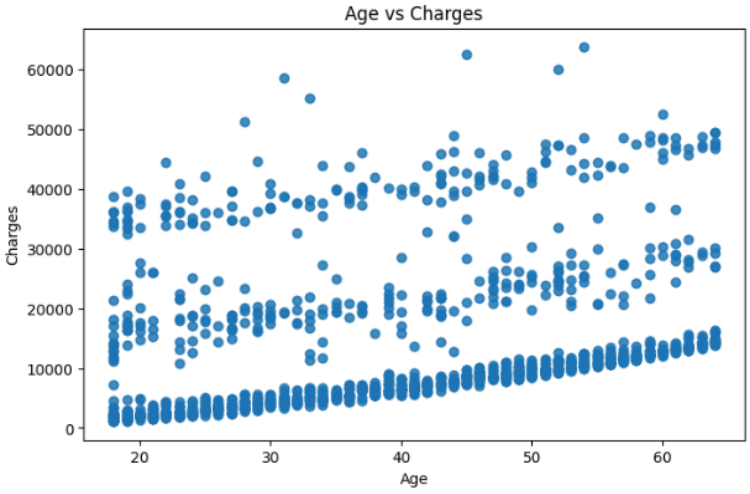
- This heatmap helps identify features that might have a strong correlation with the target variable. Looking at this graph, it tells us that age and charges seem to have the highest positive correlation out of other variables. This is then followed by charges and bmi, which is second highest but is still considered relatively weak positive correlation. Looking at the plot tells us which variables seem to be influencing charges the most, which can help with feature selection later on if needed for the models. Non numeric variables weren't displayed here, so it is also possible that the non-numeric variables might have higher importance and influence than the variables shown here.

- Scatterplot of bmi vs charges



- The plot shows us that there is a very weak trend between BMI and medical charges. It visually confirms that there is a non linear but complex relationship between BMI and charges. Based on the plot, it seems like the average BMI is around 20-35 and tends to have lower charges, there are some outliers present.

- Scatterplot of age vs charges



- It shows a clear upward trend between age and charges. As age increases, the medical charges that people tend to pay also increase. This makes sense seeing as the older someone is, the more healthcare and medical attention they might need. The scatter plot also shows us that the data appears to be clustered into 3 levels. We believe that this is telling us how other factors also play a role in determining charges alongside age. This allows us to move forward with the information that the model would need to consider multiple variables.

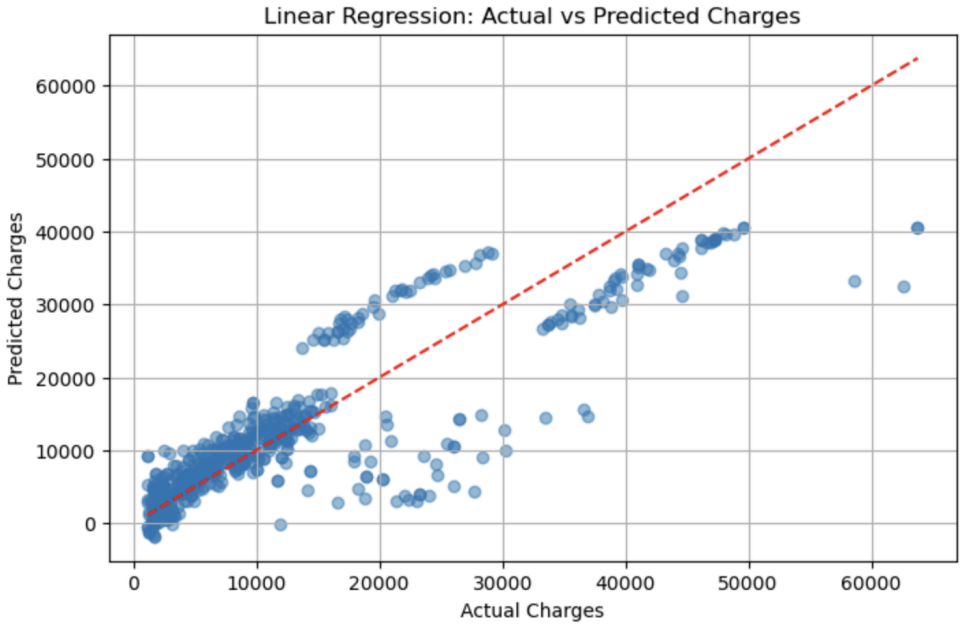
Models & Results:

1) **Linear Regression - Model Completed by: Anudeepthi Senthil Kumar**
    a) **Set up:** All categorical variables (sex, smoker, and region) performed one-hot encoding prior to model construction. Because linear regression relies on numerical inputs and is sensitive to feature scaling, this pretreatment was crucial.
    b) **Model Description:** A straightforward but effective algorithm for modeling the linear relationship between input features and the target variable (charges) is called linear regression. In order to reduce the squared discrepancy between actual and anticipated charges, it estimates coefficients for every input variable. The baseline can be this model, since it gives us a point of comparison for more intricate models.
    c) **Evaluation Metrics:**
        i) **Mean Absolute Error (MAE)**: **4,160.25**
        The average absolute difference between predicted and actual charges is about $4,160.
        This indicates moderate predictive error in dollar terms.
        ii) **Mean Squared Error (MSE)**: **39,933,194.55**
        Large squared error, showing some predictions are far off (likely due to outliers).
        iii) **Root Mean Squared Error (RMSE)**: **6,319.27**
        A more interpretable metric (same unit as charges). On average, our predictions are off by about $6,300.
        iv) **R-squared (R^2)**: **0.7398**
        The model explains approximately **74% of the variance** in the charges.
        This is reasonably good for a linear model and shows that the model captures the majority of signals in the data.
    d) **Model Interpretation:** For a straightforward linear method, the model's performance was fairly high. But the actual vs. expected values scatterplot shows:
        i) For high-cost cases, underfitting occurs (predicted values frequently underestimate)

ii) Relationships that are not linear and cannot be captured by the linear model

Therefore, more adaptable models may enhance prediction, particularly for more expensive outliers or intricate feature interactions, even though linear regression provides a solid foundation.

```
Linear Regression Performance:
MAE:  4160.25
MSE:  39933194.55
RMSE: 6319.27
R²:   0.7398
```



2) **Decision Tree Regression - Model Completed by: Netra Amin**
   a) **Set up:** The categorical variables (sex, smoker, and region) were one-hot encoded before the model training to convert them into numerical inputs which would be compatible with the decision tree regressor. The model uses a maximum depth of 5 to balance complexity and overfitting. In addition, I applied Recursive Feature Elimination for feature selection.
   b) **Model Description:** Decision Tree Regression a nonlinear model that partitions data recursively based on feature values, which makes it suitable for capturing complex relationships between the categorical and numerical variables without requiring extensive feature transformations. It works by dividing the data into subsets to reduce prediction error at each node. The feature importance metrics and RFE assist in identifying the most important predictors, providing valuable information about the predictors of insurance charges.
   c) **Evaluation Metrics:**

i) **Mean Absolute Error (MAE):** 2529.42
The average absolute difference between the predicted and actual charges is approximately $2,529. This shows relatively low average error in dollar terms compared to the typical insurance charges.

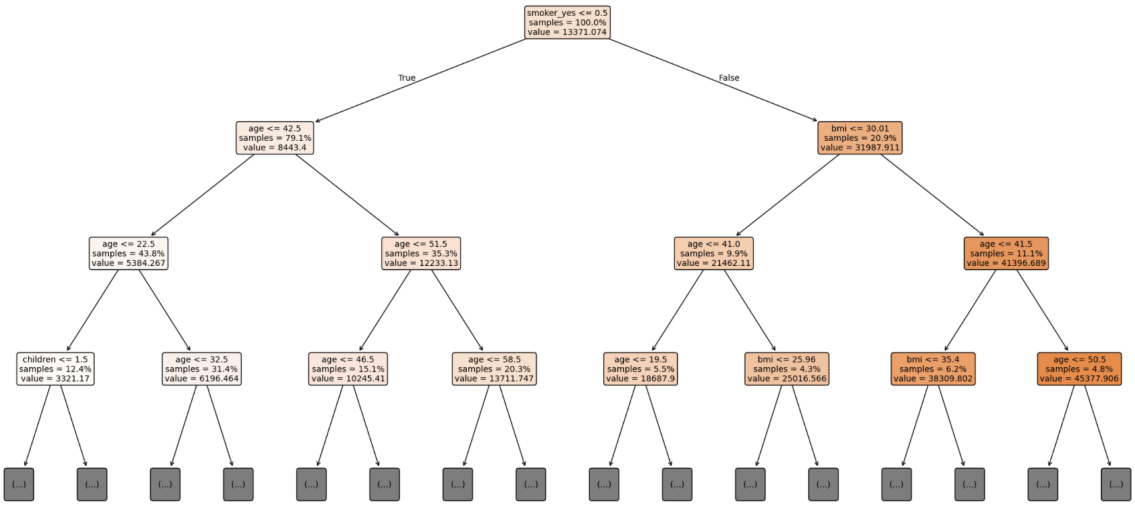
ii) **Mean Squared Error (MSE):** 21541341.60
It highlights some larger deviations between predictions and actual values which may be due to outliers or variability in charges.

iii) **Root Mean Squared Error (RMSE):** 4641.27
On average, predictions deviate by about $4,641 from the actual charges

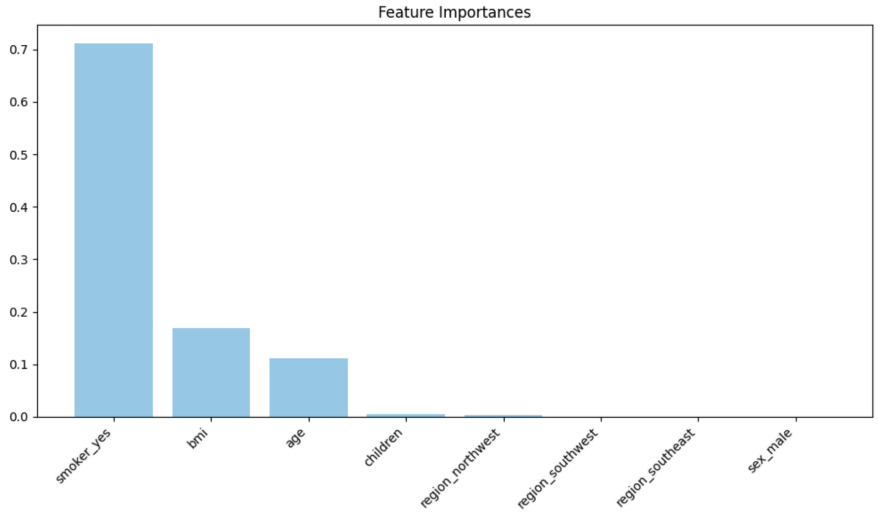
iv) **R-Squared (R^2):** 0.860
The model explains roughly 86% of the variance in insurance charges. It shows a strong predictive performance and a good fit to the data.

d) **Model Interpretation:** The feature importance ranking shows that smokers_yes, region_northwest, age, bmi, and children have the strongest influence on charges. The pruned decision tree (displaying the top three levels) shows how the model segments the data to make predictions, which improves interpretability. In addition, this method tends to capture nonlinear effects and interactions better

```
=== Evaluation Metrics ===
MAE: 2529.42
MSE: 21541341.60
RMSE: 4641.27
R²: 0.860
```



than simpler models, making it well-suited for datasets with feature types.



3) **Random Forest Regression- Model Completed by: Nanditha Chevula**
   a) **Set up:** The categorical variables (sex, smoker, and region) were one-hot encoded before the model training to convert them into numerical inputs which would be compatible with the Random Forest Regressor. The data was divided into training and testing sets using a 80/20 split. Later, feature selection was applied to improved interpretability.
   b) **Model Description:** Random Forest Regression is an ensemble learning method that constructs multiple decision trees and averages their predictions. This approach helps with reducing overfitting and capturing complex, nonlinear relationships between input variables and the target variable(charges). This model is able to handle both continuous and categorical variables(after encoding) making it super useful for our chosen dataset.
   c) **Evaluation Metrics:** I have 2 sets of evaluation metrics. This is because I created a RandomForestRegressor model with all input features and a second model with only specific features based on feature importances.
      i) **Full Feature set:**
         (1) **Mean Absolute Error (MAE): 1279.1669853384874**
            (a) The average prediction is off by about 1279.17. This is lower than the two models discussed before this.
         (2) **Mean Squared Error (MSE): 7566032.685138373**
            (a) The average squared difference between predicted and actual values is 7566032.69. This is still lower than the previous models
         (3) **Root Mean Squared Error (RMSE): 2750.642231395856**
            (a) The square root of MSE is 2759.64. It tells us that predictions only deviate by roughly 2750.

(4) **R-Squared (R^2): 0.9507037692209687**

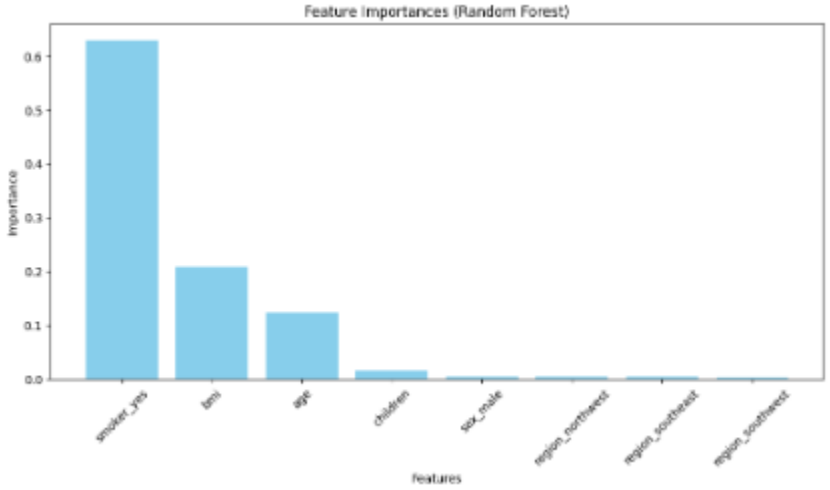
(a) The model explains about 95% of the variance in the insurance charges. This seems like a significant improvement compared to the previous two models.

d) Feature Importance and Selection:

i) Using the feature importance function, I wanted to identify the most influential features. These were the results:

(1)

	Feature	Importance
4	smoker_yes	0.629522
1	bmi	0.208603
0	age	0.124686
2	children	0.016596
3	sex_male	0.005759
5	region_northwest	0.005670
6	region_southeast	0.004955
7	region_southwest	0.004209



ii) I decided to use smoker_yes, bmi, age, and children. This is because after children, the influence from the other columns seemed negligible.

e) Evaluation Metrics(Reduced Feature Model):

i) **MAE: 1388.2712107012605**

ii) **MSE: 9072401.14486158**

iii) **RMSE: 3012.0426864275314**

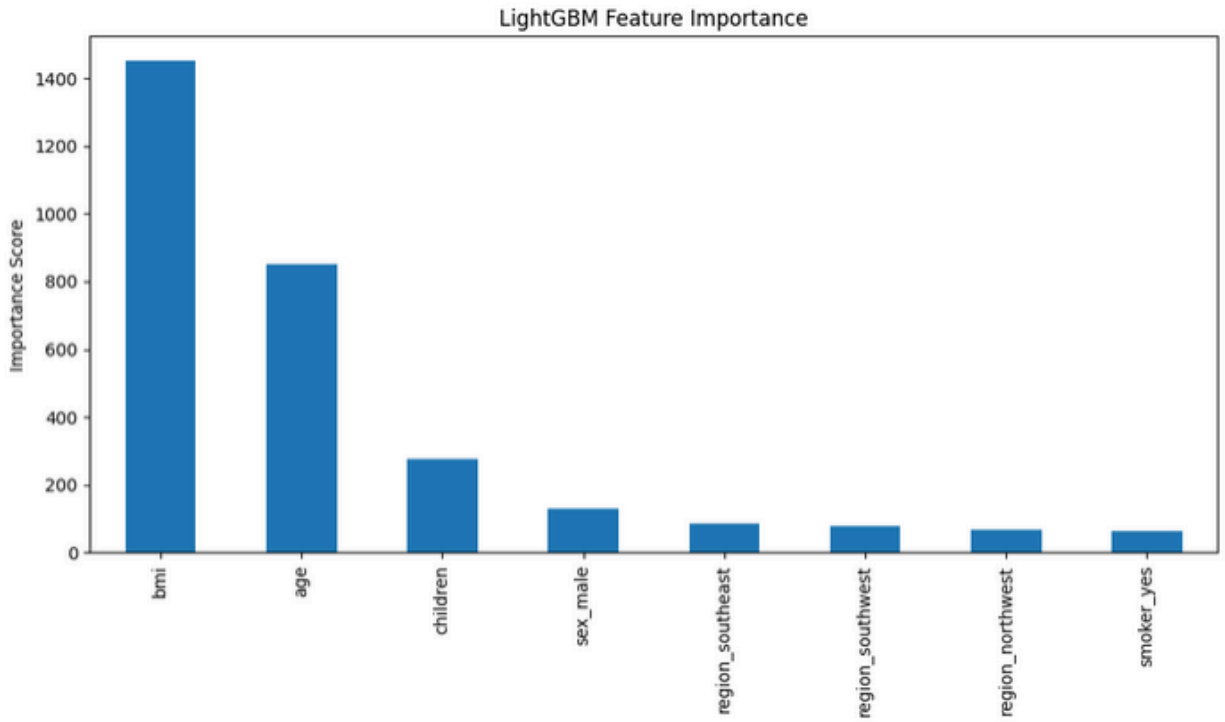
iv) **R^2: 0.9408890763272105**

v) The model still performs well, however, the overall performance was decreased slightly when compared to the full feature set model. However, this second model does help simplify and increase interpretability.

f) **Model Interpretation:** Looking at the performance on an overall scale, the Random Forest model outperformed the previous 2 models on all evaluation metrics, both full feature and simplified. It has a relatively high performance and

was able to capture the different nonlinear patterns and feature interactions that we noticed during the exploratory data analysis.

4) **Gradient Boosting Regression (LightGBM) - Model Completed by: Yusuf Kshem**
    a) **Set up:** Categorical variables (sex, smoker, and region) were transformed using one-hot encoding prior to model construction. The charges column was set as the target, and the remaining columns were used as features. The dataset was split into training and testing sets with an 80/20 ratio. Unlike linear models, normalization is not strictly necessary with tree-based models like LightGBM.
    b) **Model Description:** LightGBM (Light Gradient Boosting Machine) is a high-performance, tree-based boosting algorithm that builds models sequentially. Each tree tries to correct the prediction errors of the previous one. This model is known for its efficiency and accuracy, especially on structured datasets. The model was trained with default parameters using the training set, and predictions were made on the test set. Feature importance was extracted directly from the model to analyze which variables contributed most to the prediction accuracy.
    c) **Evaluation Metrics:**
        i) **Mean Absolute Error (MAE)**: **2,166.43**
        The average absolute difference between predicted and actual charges is approximately $2,166.
        ii) **Mean Squared Error (MSE)**: **14,737,041.03**
        Indicates the presence of some larger errors; penalizes these more than MAE.
        iii) **Root Mean Squared Error (RMSE)**: **3,838.89**
        On average, the model's predictions deviate from the true charges by about $3,839.
        iv) **R-squared (R^2)**: **0.9040**
        The model explains approximately 90.4% of the variance in insurance charges, which shows strong performance and good generalization.
    d) **Model Interpretation:** The LightGBM model performed competitively, with strong accuracy and the second-lowest MAE among the models tested. It effectively captures nonlinear relationships and complex feature interactions without overfitting. The performance suggests that LightGBM is a strong candidate for insurance charge prediction.



5) Conclusion

a) Overall, all four models performed reasonably well in predicting our target variable. Linear Regression provided a strong baseline, explaining about 74% of the variance, but it did struggle to capture the complex and nonlinear relationships between features. The Decision Tree Regression performed better, explaining approximately 86% of the variance, and was able to model non-linear connections that the linear model could not. LightGBM demonstrated robust accuracy and strong generalization. Among all approaches, the ensemble models (Random Forest Regression and LightGBM) performed the best, capturing more variance and yielding lower MAE values. These models excelled due to their ability to handle complex feature interaction and maintain robustness against outliers. Out of all the models, Random Forest Regression had the most favorable evaluation metrics, even after a slight decrease in performance after the feature selection. While the ensemble models already perform well, there is still room for improvement. In the future, utilizing better feature selection methodologies as well as encoding methods, we could achieve an even better accuracy and further reduce the risks of overfitting or underfitting with our models.