Project Report: Predicting Medical Insurance Cost Using Machine Learning

Project Overview

The goal of this project is to predict medical insurance costs for individuals using machine learning algorithms. The predictions are based on demographic and lifestyle factors such as age, gender, BMI, smoking status, number of children, and residential region. Accurate predictions of insurance costs help:

- Insurance companies create better pricing strategies.
- Reduce the risk of underpricing or overpricing policies.
- Encourage individuals to adopt healthier lifestyles to reduce premiums.

Problem Statement

Medical insurance costs are influenced by multiple factors, which interact in complex ways. Traditional linear methods may not capture these relationships effectively.

Challenge: Build a predictive model using machine learning to accurately estimate insurance charges, considering features like smoking habits, BMI, age, and location.

Objectives

The objectives of the project are:

1. Perform Exploratory Data Analysis (EDA) to understand the dataset.
2. Preprocess the data to make it suitable for machine learning models.
3. Build and evaluate multiple regression models:
   - Linear Regression
   - Random Forest Regressor
   - XGBoost Regressor
4. Analyze the performance of each model using evaluation metrics (R^2, RMSE, and MAE).
5. Identify the most influential features affecting insurance costs.
6. Provide actionable insights for insurance companies based on model results.

Dataset Overview

The dataset contains 1338 rows and 7 features, with the target variable being 'charges'.

Features:

- age: Age of the individual (numeric)

- sex: Gender: male/female (categorical)
- bmi: Body Mass Index (numeric)
- children: Number of dependents (numeric)
- smoker: Smoking status: yes/no (categorical)
- region: Residential region (categorical)
- charges: Insurance cost (target variable).

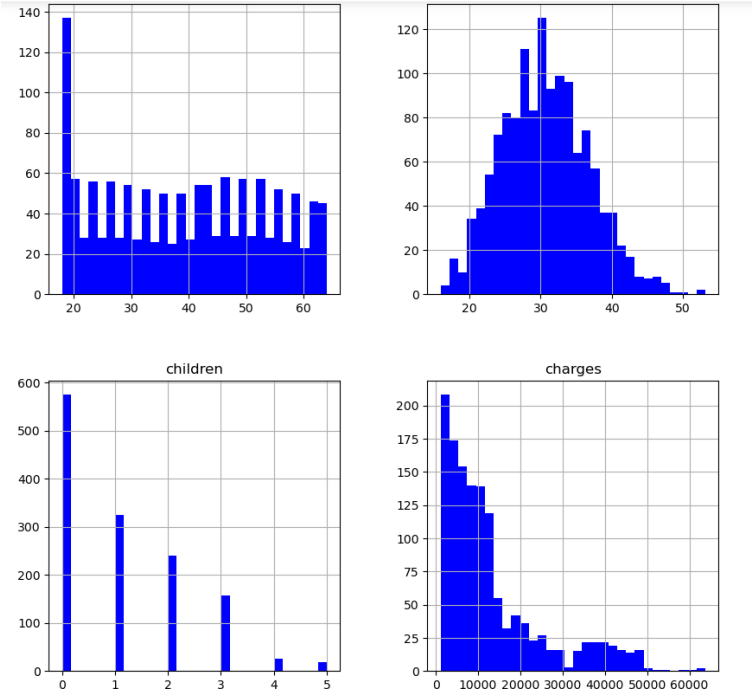
Exploratory Data Analysis (EDA)

1. Distribution of Features:
- Age: Most individuals are aged between 18-65.
- BMI: Normally distributed with an average of ~30.
- Charges: Right-skewed, indicating few high-cost individuals.

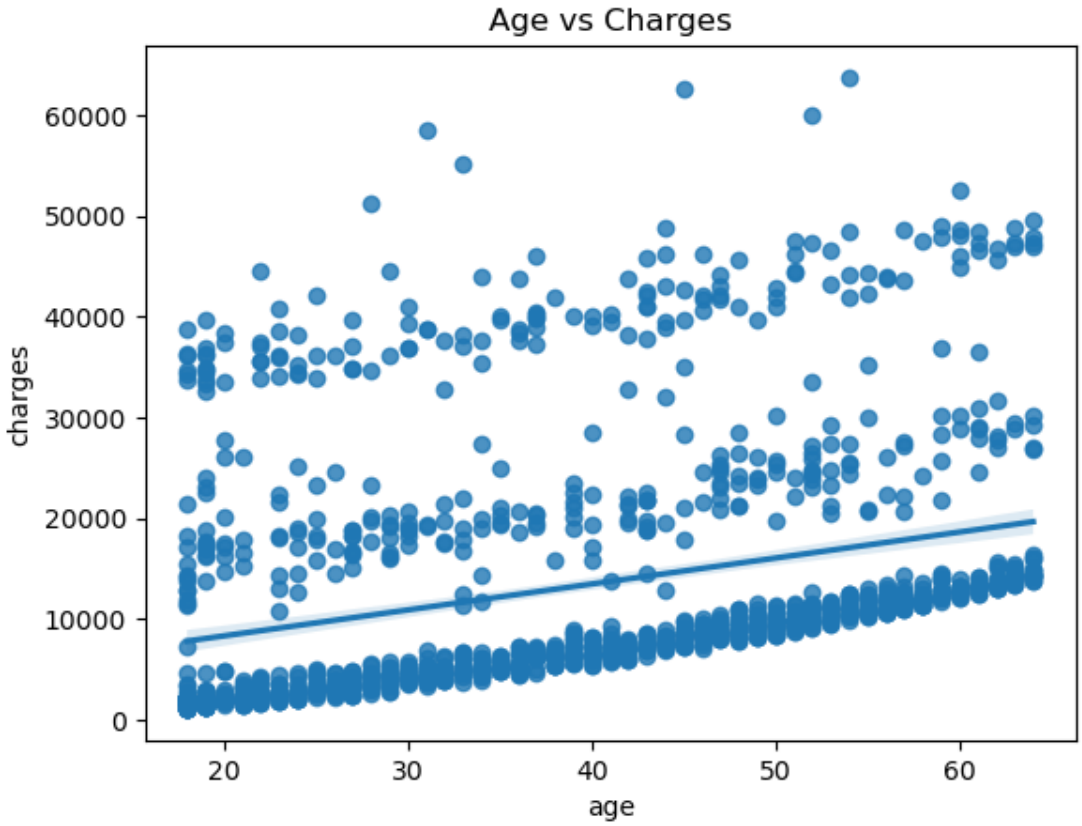
2. Relationships with Target Variable:
- Age vs Charges: Costs increase with age.
- BMI vs Charges: Higher BMI leads to higher costs, especially for smokers.
- Smoker Status: Smokers incur the highest costs.

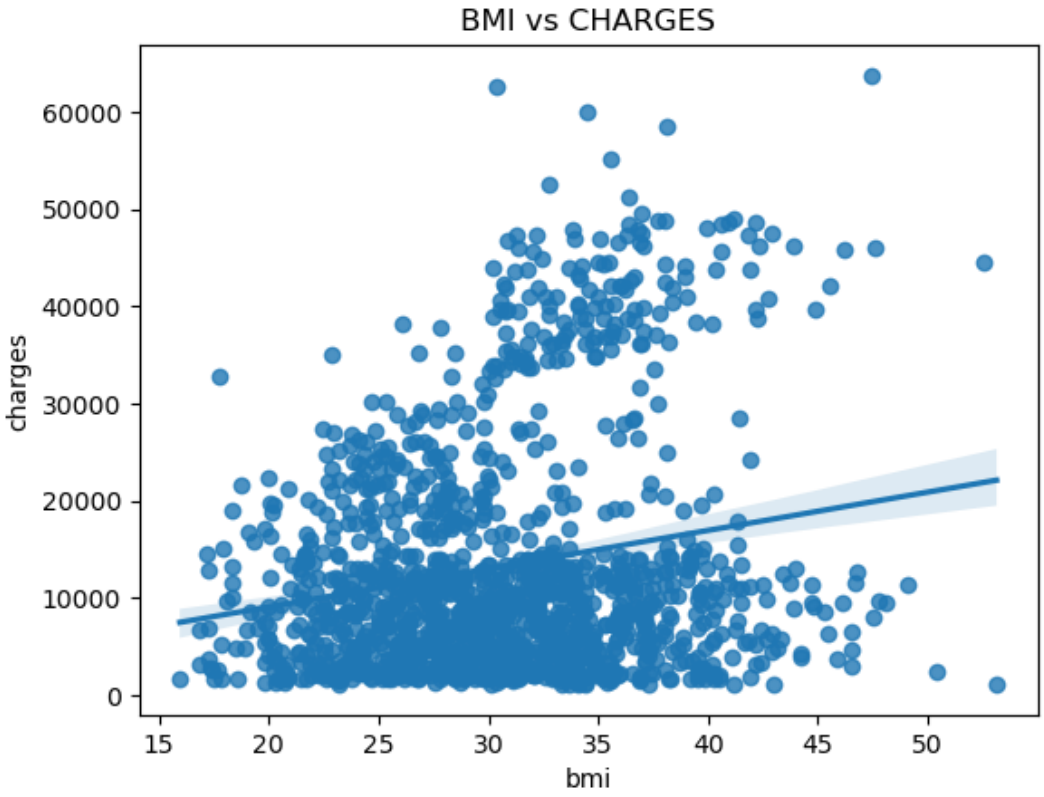
The correlation heatmap showed that the `smoker` variable has the strongest positive correlation with `charges`.

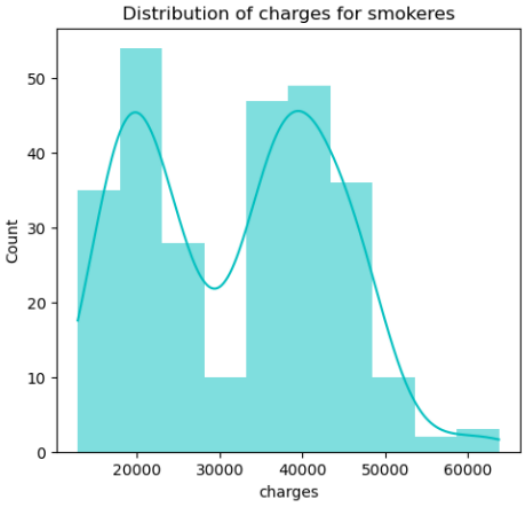
Distribution of Features

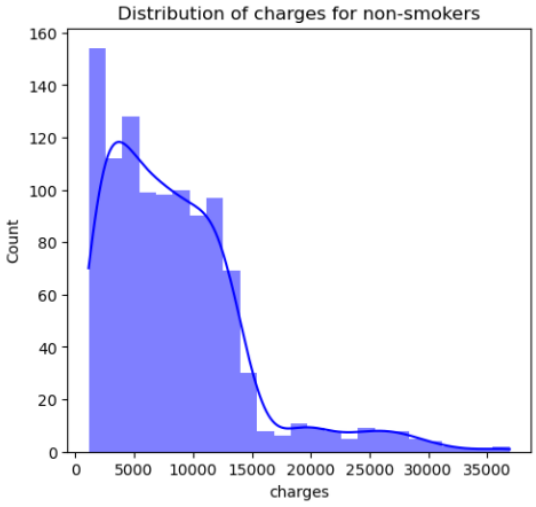


Relationships Between Features and Charges

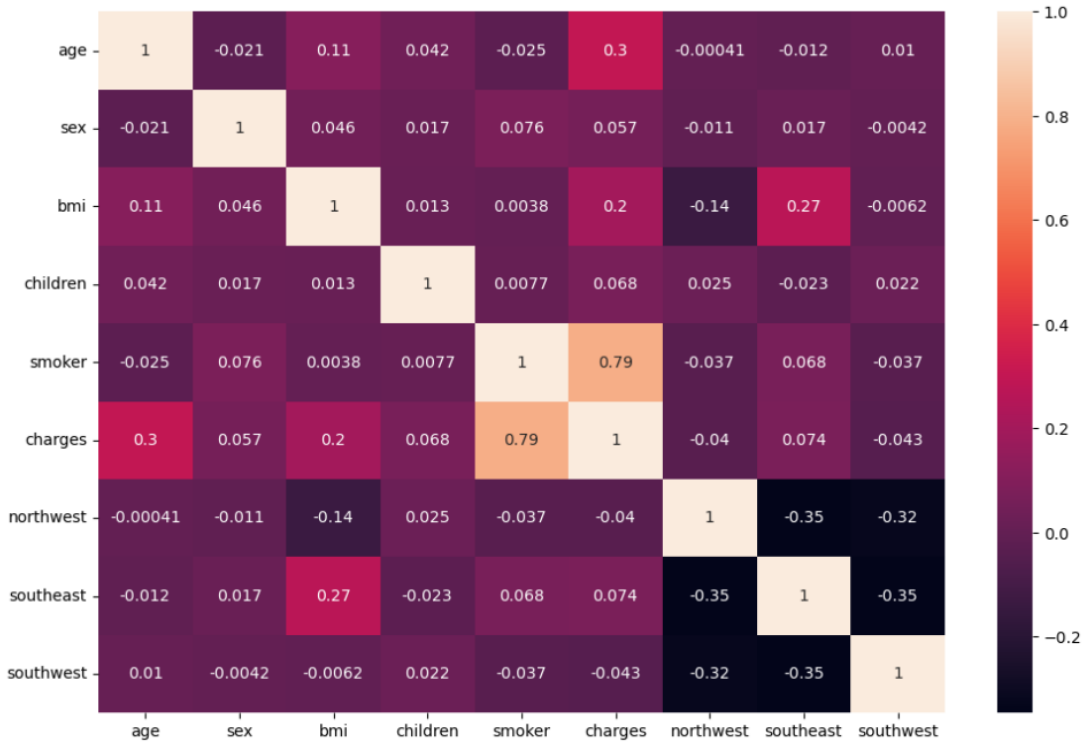



BMI vs CHARGES
charges
bmi
60000
50000
40000
30000
20000
10000
0
15
20
25
30
35
40
45
50

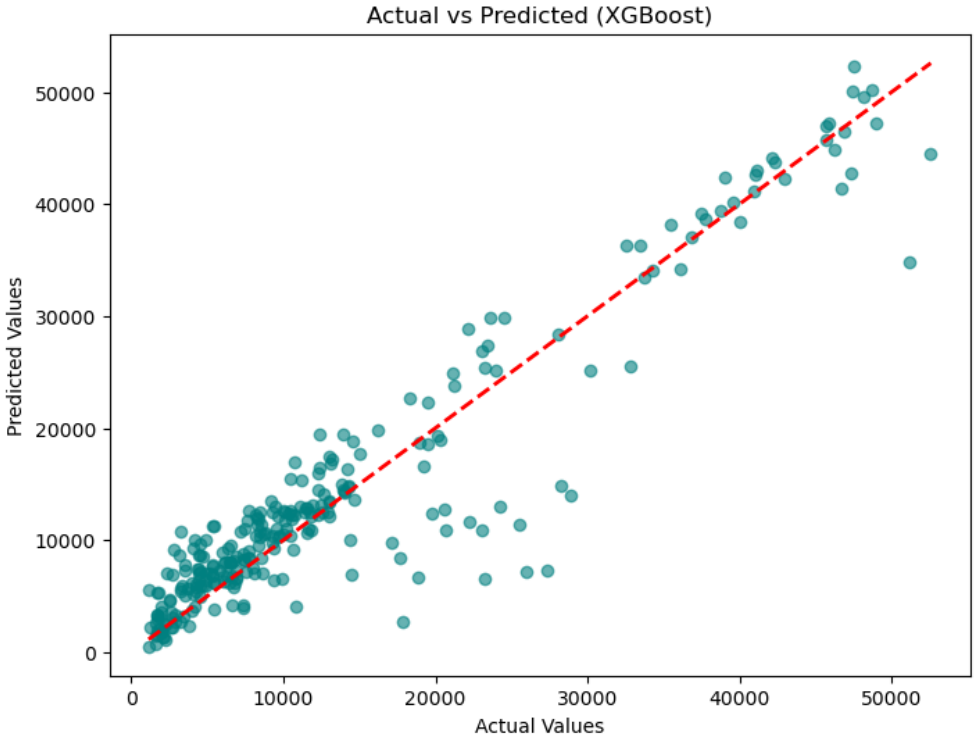

Distribution of charges for smokeres
Count
charges
50
40
30
20
10
0
20000
30000
40000
50000
60000

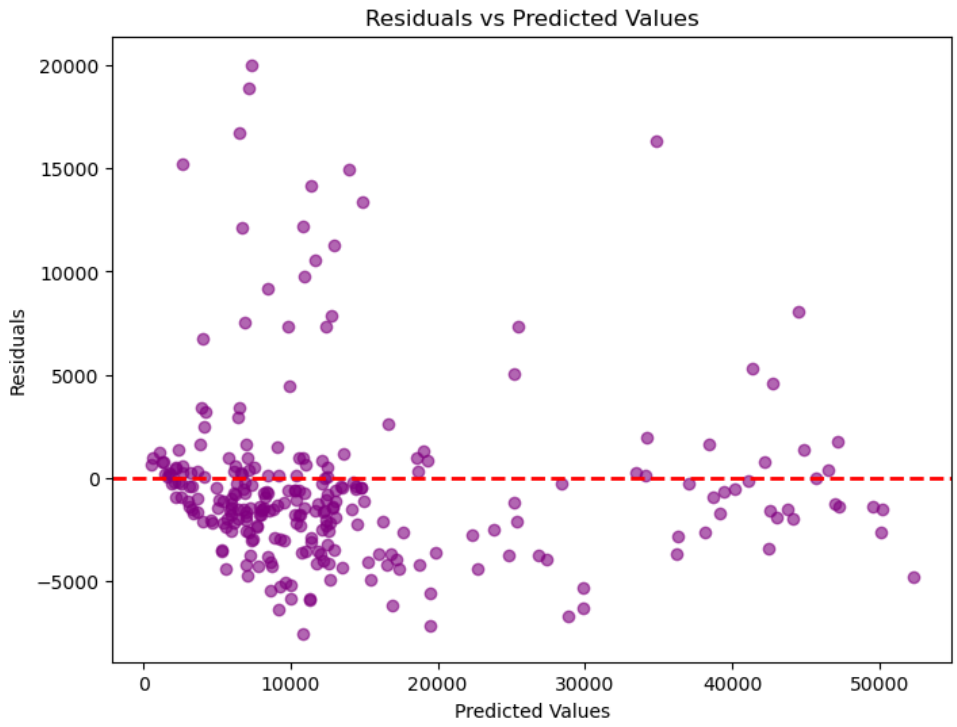

Distribution of charges for non-smokers
Count
charges
160
140
120
100
80
60
40
20
0
0
5000
10000
15000
20000
25000
30000
35000

Correlation Heatmap

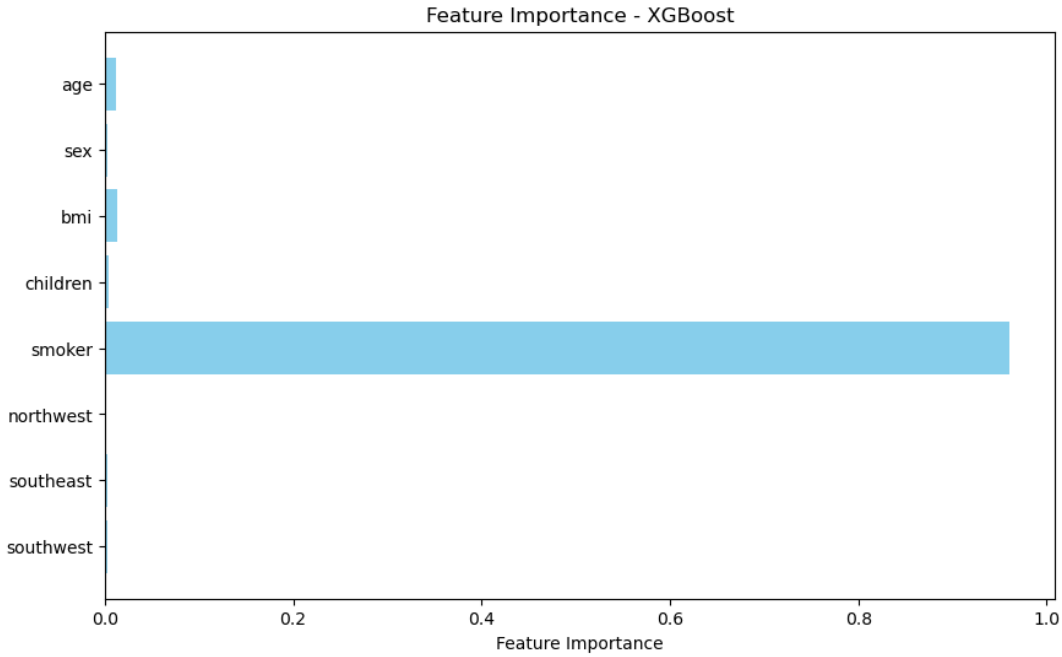


Model Evaluation





Feature Importance



Conclusion

The project successfully developed a machine learning model to predict medical insurance costs. The **XGBoost Regressor** provided the best results with an R^2 score of **0.92**.

Key findings:
- Smoking status is the most significant predictor of costs.
- BMI and age also play critical roles.

These insights enable insurance companies to optimize pricing strategies and promote healthier behaviors among customers.