Analyzing Factors Influencing Insurance Cost

—

Jameson Tuck

Problem Statement

The goal of this analysis is to investigate the factors influencing insurance costs (charges) and quantify their impact. Specifically, I aim to assess how demographic and lifestyle variables, such as age, sex, BMI, smoker status, number of children, and region, contribute to variations in insurance charges.

This analysis will focus on identifying relationships between these predictor variables and insurance costs to provide insights into their relative importance and potential predictive power.

Key Hypotheses

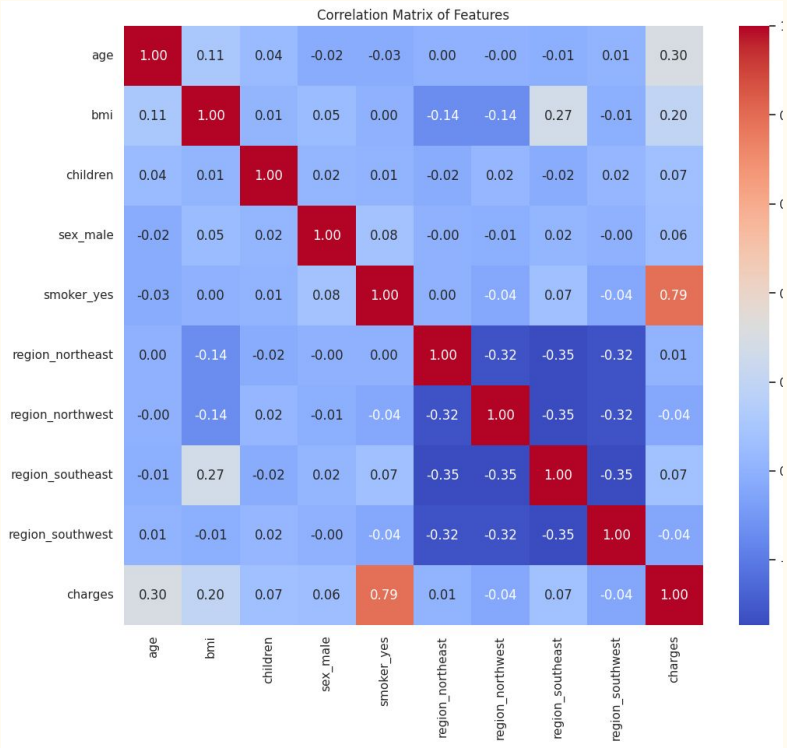
- Insurance costs are higher for individuals who smoke.
- Individuals with a lower BMI incur lower insurance costs.
- Having children increases insurance costs.
- Older individuals face higher insurance costs.

Data Collection, Cleaning, and Preparation

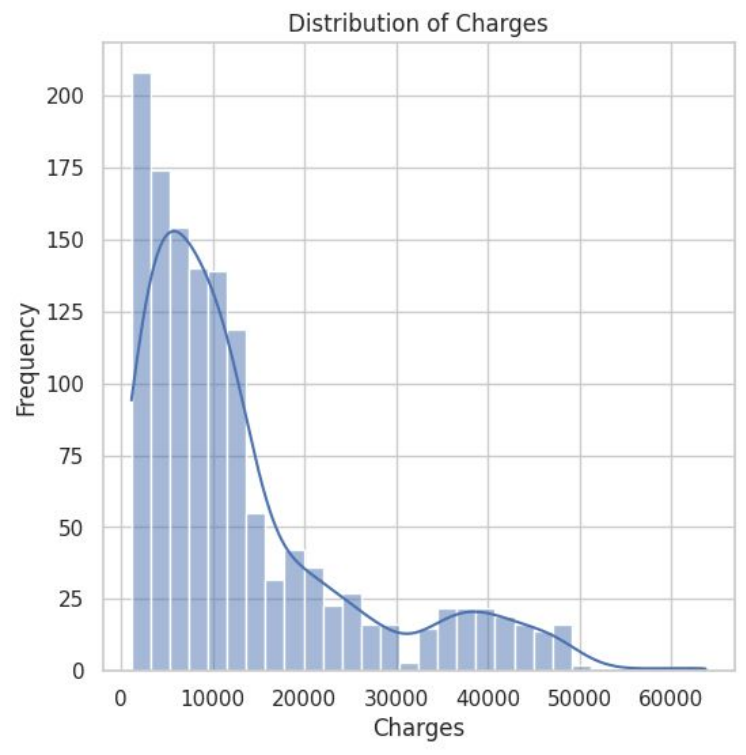
- The data was collected and loaded using Kaggle
- The collected data was then cleaned by inspecting the shape of the dataframe and dropping rows with any missing values
- The categorical variables were transformed using OneHotEncoder and the original columns were dropped

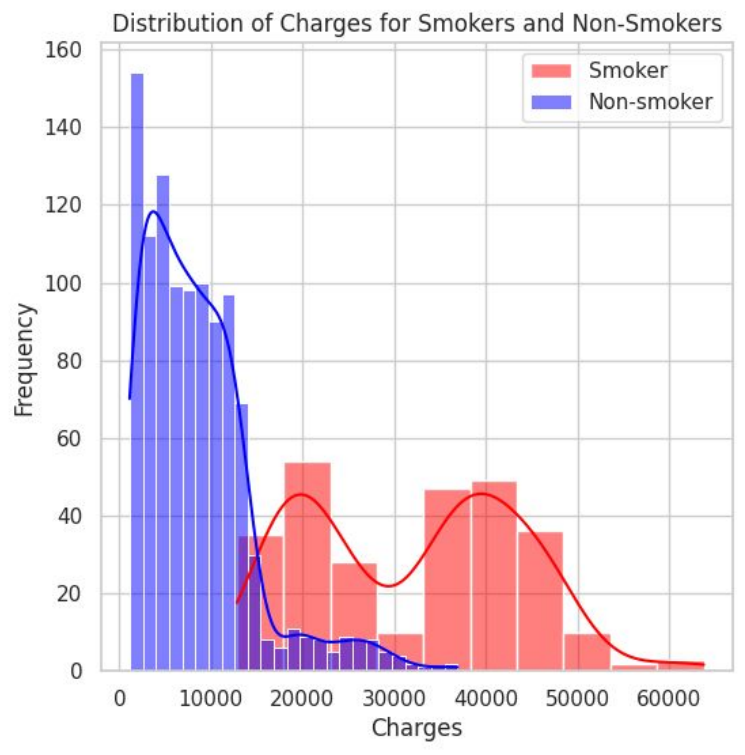
Exploratory Data Analysis (EDA)

- High linear correlation between smoking and charges
- Weak linear correlation between BMI and age to charges
- No linear correlation between children and charges

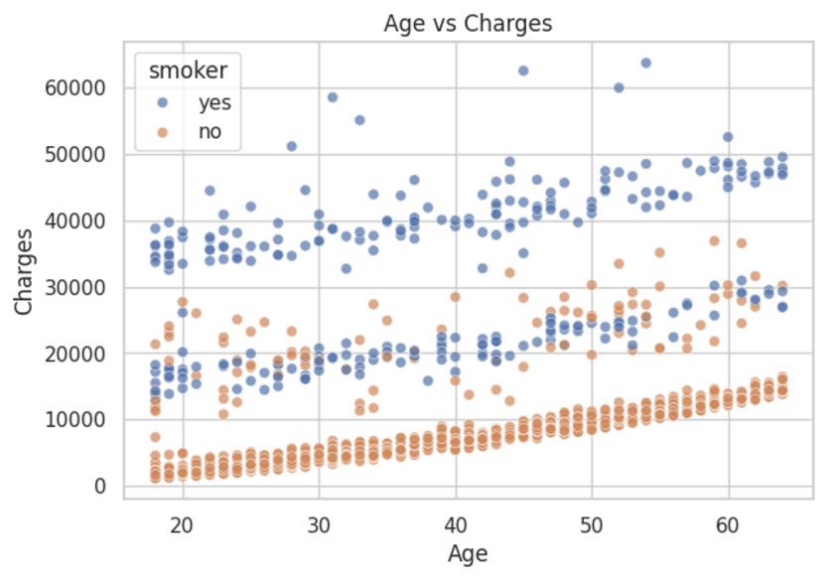


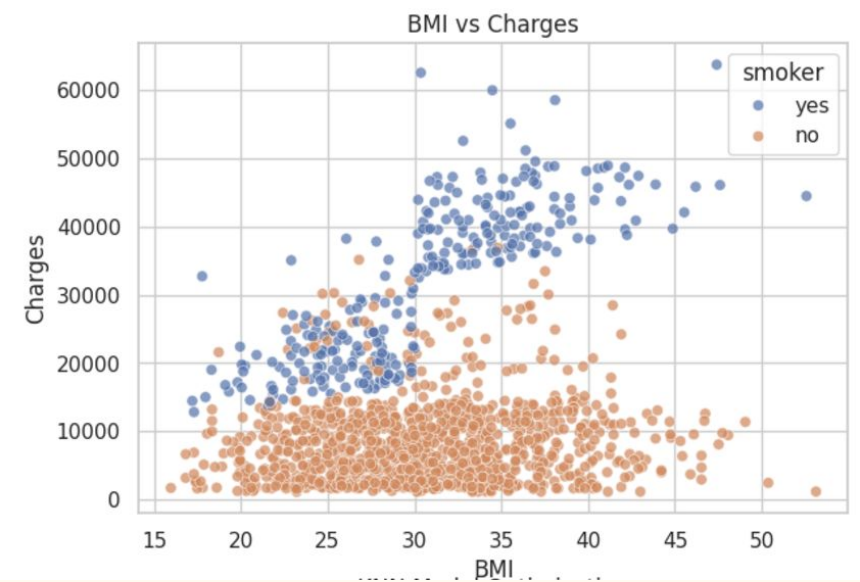
Exploratory Data Analysis (EDA) and Visualization





Continued Visualization





Models and Regularization Techniques

- Linear regression with L1 (LASSO) regularization
- Decision Tree with GridSearchCV - performs grid search across a range of hyperparameters
- Random Forest with with GridSearchCV to tune hyperparameters (ie. max depth and split)
- K-nearest neighbors with standardized data and optimal K value
- Dense neural network with dropout layers

Results

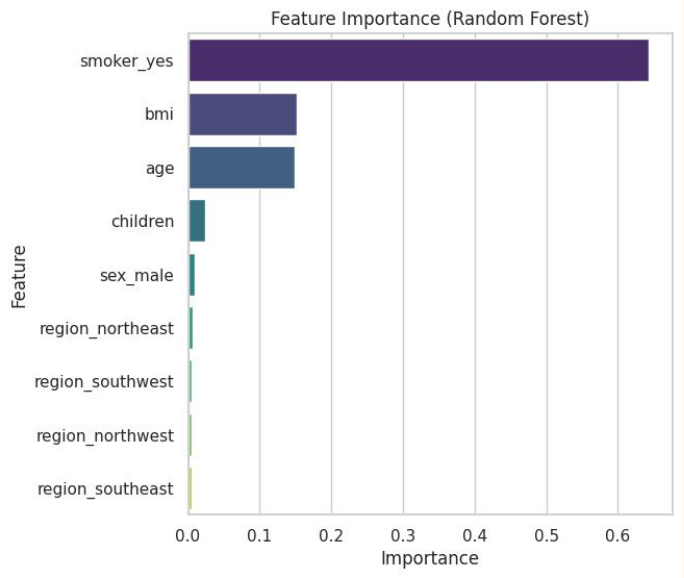
Model	Train MSE	Test MSE
Linear Regression	37,277,711	33,605,932
Decision Tree	21,120,357	22,812,669
Random Forest	8,273,732	20,155,211
Dense Neural Network	34,704,355	30,746,318
K-Nearest Neighbor	18,119,239	29,058,240

Analysis

- Even though the MSEs are quite large, we are working with large numbers so this is natural.
- The values still seem quite large when you take the root of them, but this is because of outlier data points and in most cases, the model generalizes to a few hundred dollars off the actual value.
- The best performing model was the random forest model.
    - The Test MSE for this model is significantly higher than the Train MSE implying that the model is overfit, however it still performs best comparatively

Conclusion

- Key Insights
  - Best Model: The Random Forest model outperformed other models, achieving the lowest test MSE.
  - Key Predictors: Analysis of Random Forest feature importance revealed that smoker status, bmi, and age are the most influential variables in predicting insurance charges.
- Limitations
  - Overfitting in the Random Forest model indicates a need for further regularization or additional training data.
- Future Directions
  - Incorporate additional predictors, such as exercise habits or family medical history, to improve predictive power.



References

https://www.kaggle.com/datasets/mirichoi0218/insurance/data

Thank you!