SDS 2025: MISTERS

By Danvern, James, Julian

PROBLEM FRAMING & OBJECTIVE

Problem Framing

- We will be using medical insurance charges as our target variable and age, sex, BMI, children, smoker and region as our feature variables
- A linear regression model will be used as our baseline model and an XGBoost model will be used as our main model
- Exploring feature importance using the XGBoost model

Objective:

- Predicting insurance charges based on 6 different feature variables namely, age, sex, BMI, children, smoker and region.
- Providing insights on the top factors affecting insurance charges.
- Assessing whether insurance charges vary across sex and region
- Providing recommendations to insurance companies based on our findings

EXPLORATORY DATA ANALYSIS

Exploring the relationship between medical charges and the 6 feature variables

- As age increases, charges increase
- Each age group has 3 different groups of charges

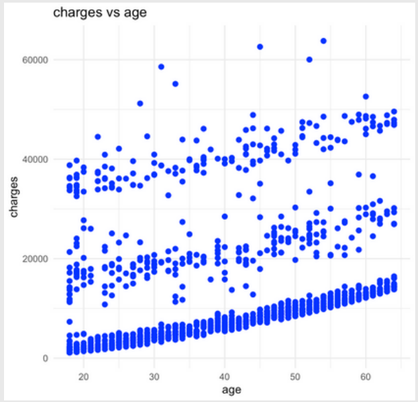


Figure 3.1. charges vs age

- Median is similar across sex
- IQR for males is slightly larger

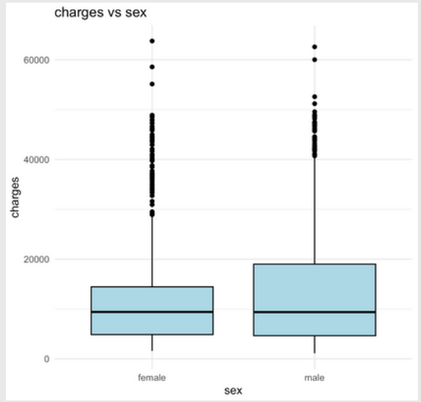


Figure 3.2. charges vs sex

- As BMI increases, charges increase
- Points with moderate charges have below median BMI
- Points with high charges have above median BMI

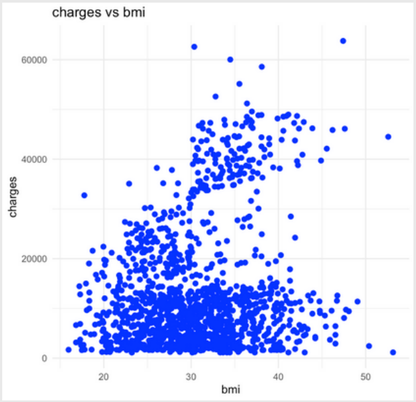


Figure 3.3. charges vs BMI

- Median varies slightly across children
- IQR varies slightly across children

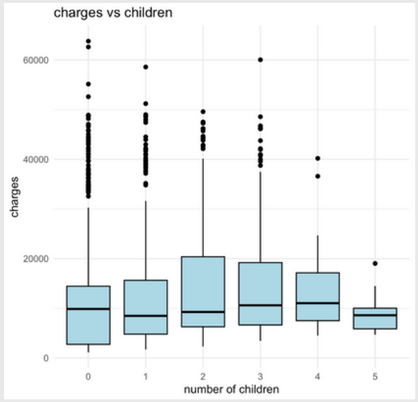


Figure 3.4. charges vs children

- Median for smokers is significantly higher
- IQR for smokers is significantly higher

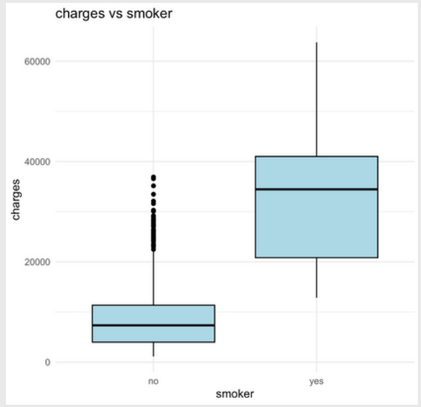


FIgure 3.5. charges vs smoker

- Median is similar across region
- IQR varies slightly with region

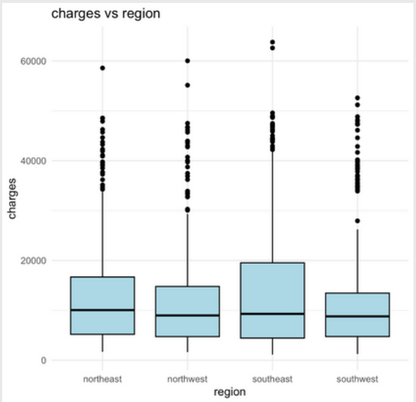


FIgure 3.6. charges vs region

EXPLORATORY DATA ANALYSIS (EDA)

INITIAL DATA SET:

- 1338 observations & 7 variables

REMOVAL OF DUPLICATES:

- Two observations are identical across all seven variables
- It is highly unlikely for two individuals to have all 7 variables being identical
- The two observations are most likely duplicates
- 1 observation was removed as retaining it could introduce bias or skew the data more than excluding it would

REMOVAL OF FEATURES:

- Conducted backward elimination
- Best model contains 6 variable; excluding sex as a feature variable

PERFORMING ONE-HOT ENCODING:

- Performed on region variable
- Avoids introducing false ordinal relationship between the different regions

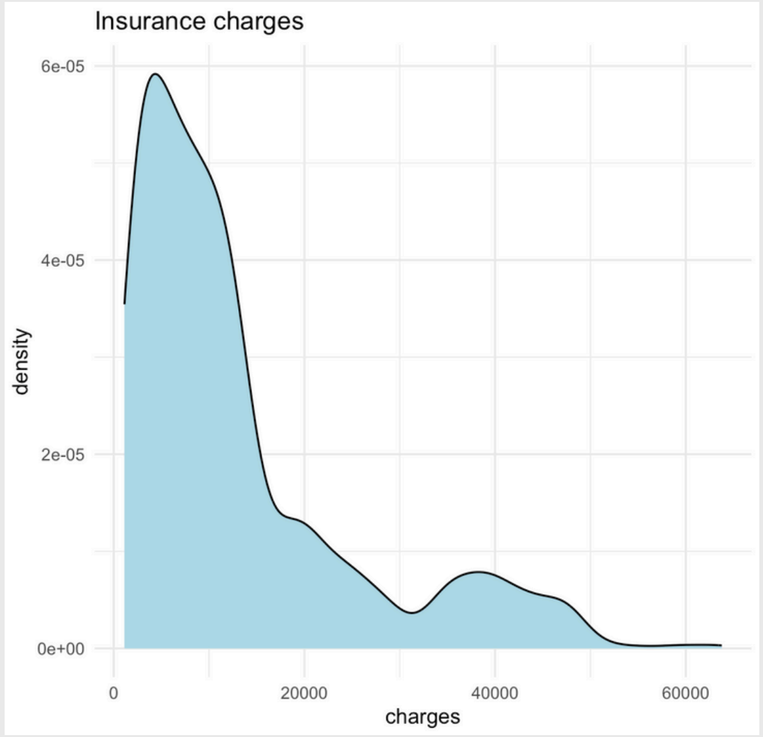
FINAL DATASET:

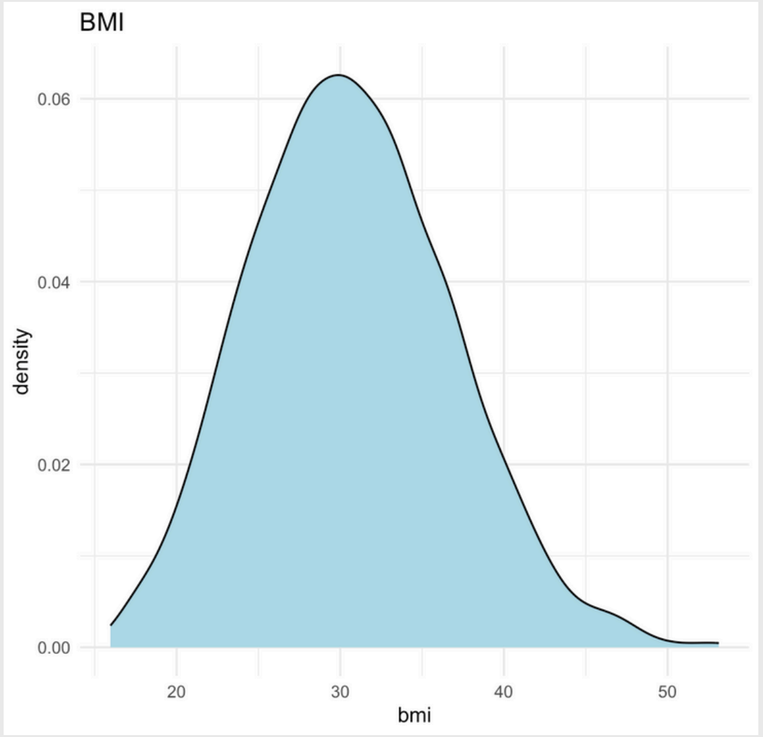
- 1337 observations and 9 variables

EXPLORATORY DATA ANALYSIS (EDA)

REMOVAL OF OUTLIERS:

- Insurance charges vary significantly depending on factors such as type of incident, severity and coverage limits.
- It is impossible to exclude people with high BMI
- Retaining these values allows the model to accurately reflect the full range of potential charges and BMI





REGRESSION AND MODELLING APPROACH

POINT OF COMPARISON:

- Root Mean Square Error (RMSE) is used to decide the accuracy of our model
- RMSE is calculated by taking the square root of the average of the squared differences between predicted and actual values
- A lower RMSE indicates a more accurate model

BASELINE:

- Multiple linear regression model is used due to its simplicity
- Predicts insurance charges by using an equation to estimate a linear relationship between the insurance charges and our 8 feature variables

ADVANCED:

- XGBoost model is used due to higher predictive accuracy and speed
- Higher processing speed allows it to be implemented on larger data sets in the future
- Predicts insurance charges by using various decision tree models, with each new decision tree model added to reduce the error in the previous model
- Decision tree model makes use of a series of yes/no questions to categorise the data based on the 8 feature variables.

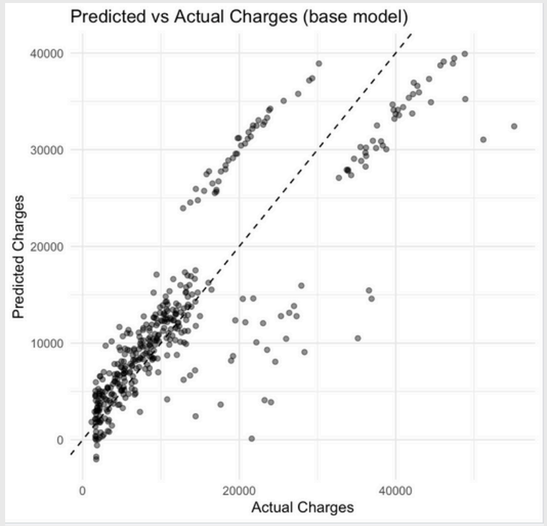
KEY FINDINGS AND VISUALISATIONS

VISUALISATION OF MODEL:

- Scatter plot of predicted vs actual medical charges to visualise accuracy
- Dotted line represents a straight line with a gradient of 1
- The more accurate the model, the closer the points are to the line
- If points lie under the line, this signifies an under-prediction and vice versa

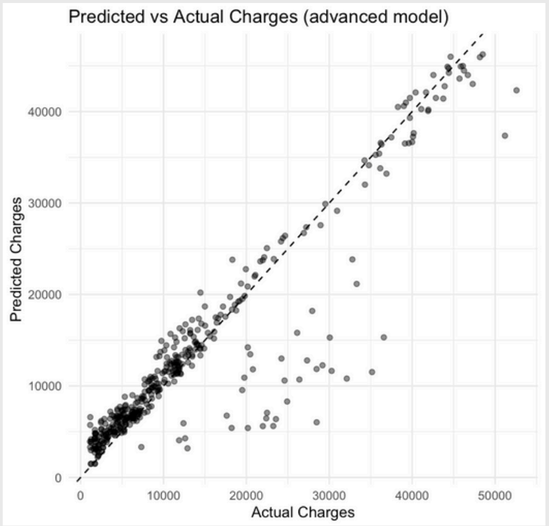
BASELINE:

- RMSE of 6015.2
- Only points with low charges lie near the line
- Majority of points with moderate charges lie above the line
- Points with high charges lie below the line



ADVANCED:

- RMSE of 4624.3
- Majority of the points lie near the line
- Few points lie significantly below the line



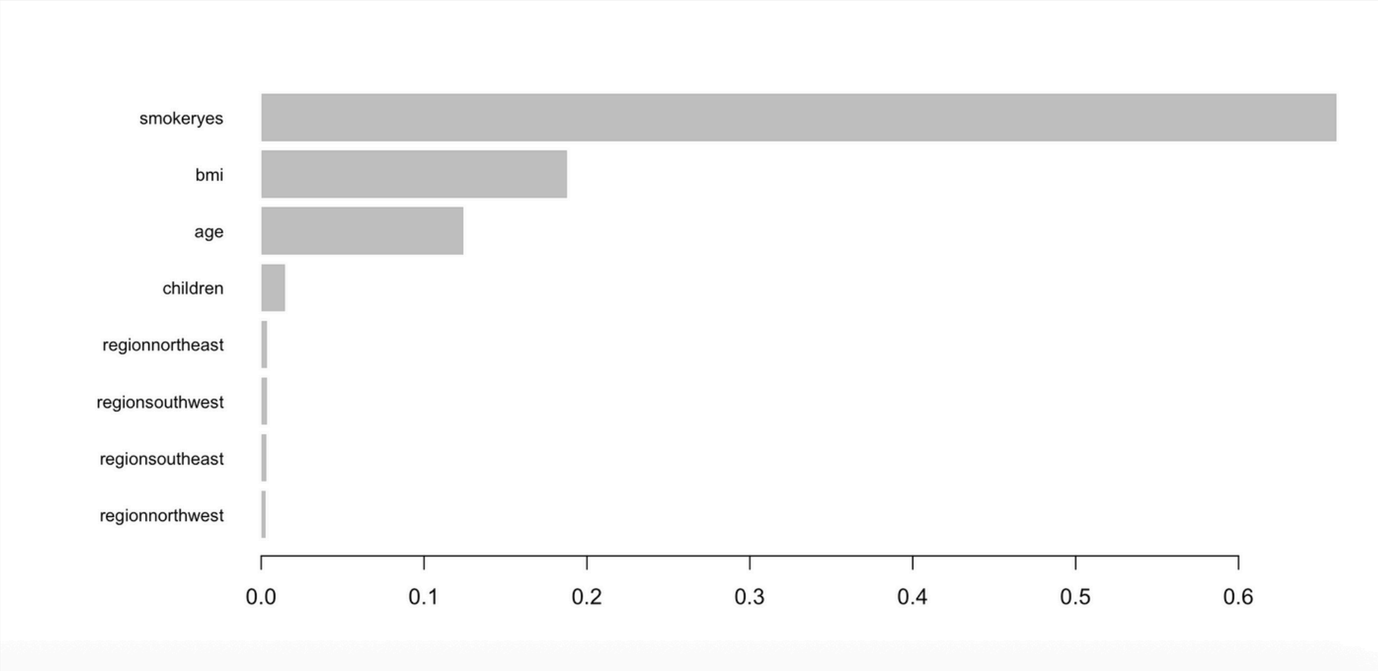
FEATURE IMPACT

XGBOOST FEATURE IMPORTANCE

XGBoost feature importance quantifies relative contribution of each feature variable to the predictions in charges.

Analysis indicates smoking status is the most significant factor influencing insurance costs, followed by BMI, age, number of children, and region.

This indicates that lifestyle choices of the individual dominate charges more than demographic.



FAIRNESS ANALYSIS

STATISTICAL TEST:

- Differences in charges due to sex is not statistically significant (p-value = 0.698), indicating fairness in charges
- Kruskal-Wallis test is used to determine differences in charges due to region
- A lower p-value indicates a more significant difference
- We will be using a significant level of 5% to determine fairness
- A p-value less than 0.05 indicates a significant difference in charges across region indicating unfairness

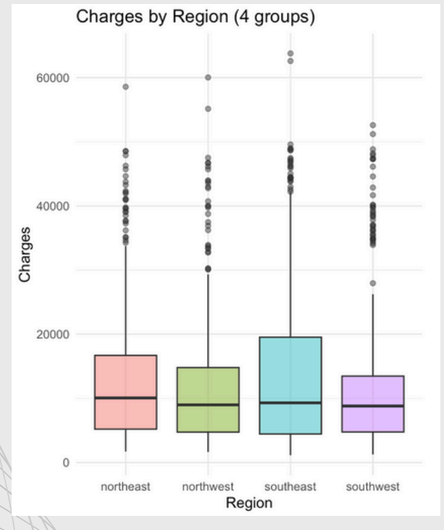


Figure 8.1. p -value = 0.192 showing no significant difference between the 4 regions

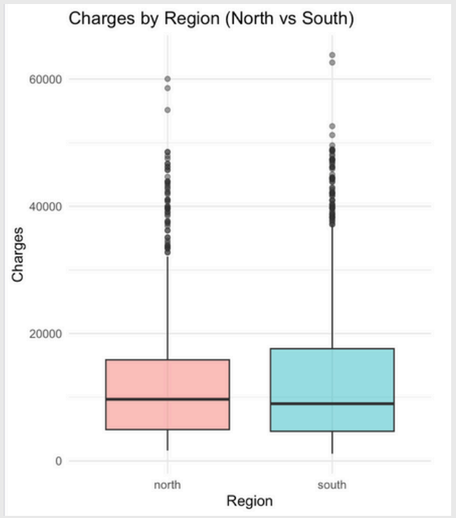


Figure 8.2. p -value = 0.443 showing no significant difference between North and South region

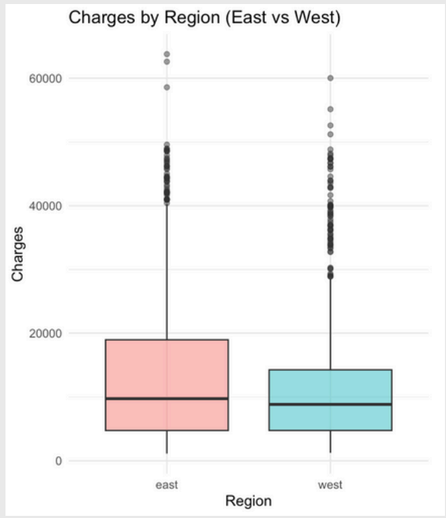


Figure 8.3. p -value = 0.0447 showing significant difference between East and West region

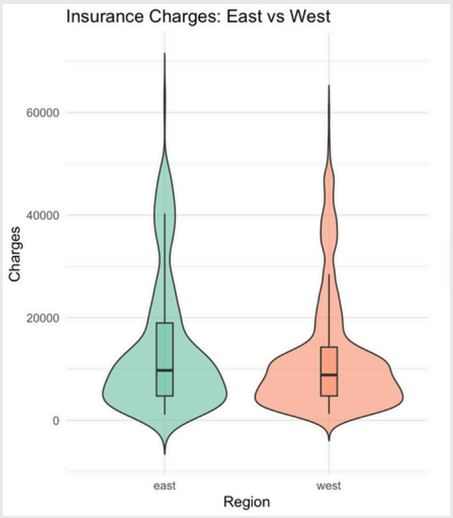


Figure 8.4. violin plot shows charges is more spread out upwards for the east while the west has more consistent and lower charges

Practical Recommendations:

FORECAST OVERALL CHARGES:

- Use the model to predict total expected insurance claims or costs over a period
- Assist companies with budgeting, risk management and resource allocation

PERSONALISED PREMIUMS:

- Use the model's predictions for individuals based on their risk profiles
- Supports pricing strategies that are competitive and financially sustainable

FEATURE INSIGHTS/ DECISION SUPPORT:

- Model can highlight which factors influence charges the most
- Guide marketing, policy design and claims management strategies.

Difficulties faced and methods used to overcome:

PROBLEM:

- Data set had limited number of feature variables making it not fully capture the relationships influencing our target variable reducing predictive accuracy

SOLUTION:

- Expand feature variables in future data sets by incorporating additional relevant features (e.g presence of chronic conditions, no. of previous claims)

PROBLEM:

- Small number of observations hence, may not be representative of the larger population

SOLUTION:

- Expand on data set to strengthen the model's statistical reliability, reducing test error

APPENDIX

https://github.com/dphyy/SDS