Medical Insurance

1.1 Overview

- There are 1,338 samples
- Each sample has 7 attributes which are given below
- There is no null value in the dataset

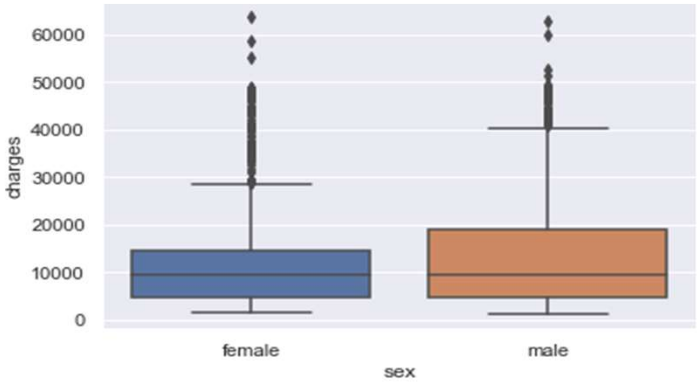
Attribute	Description	Type
Age	Age of the primary beneficiary	Integer
Sex	Policy holder's gender	Object
BMI	Body mass index (ideal BMI is within the range of 18.5 to 24.9)	Float
Children	Number of children / dependents covered by the insurance plan	Integer
Smoker	Yes or No depending on whether the insured regularly smokes tobacco.	Object
Region	Place of residence in the U.S., divided into four geographic regions - northeast, southeast, southwest, or northwest	Object
Charges	Individual medical costs billed to health insurance	Float

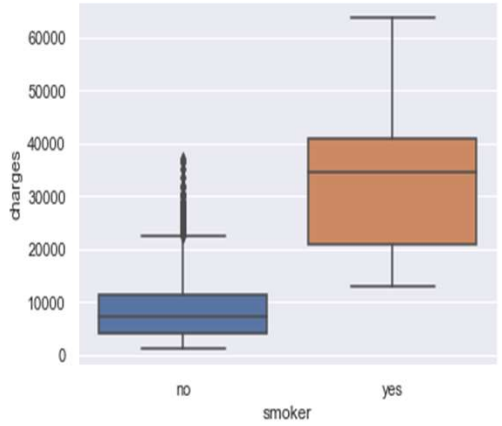
Objectives

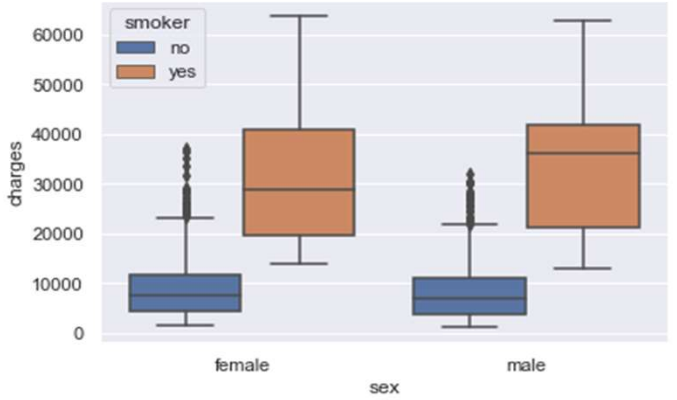
Create ML Models to:

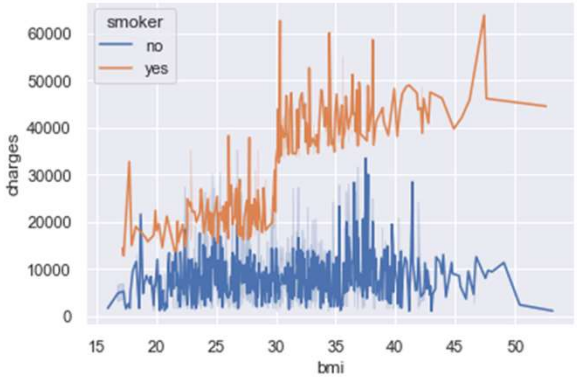
1. Predict medical claim
2. Predict whether a customer is a smoker

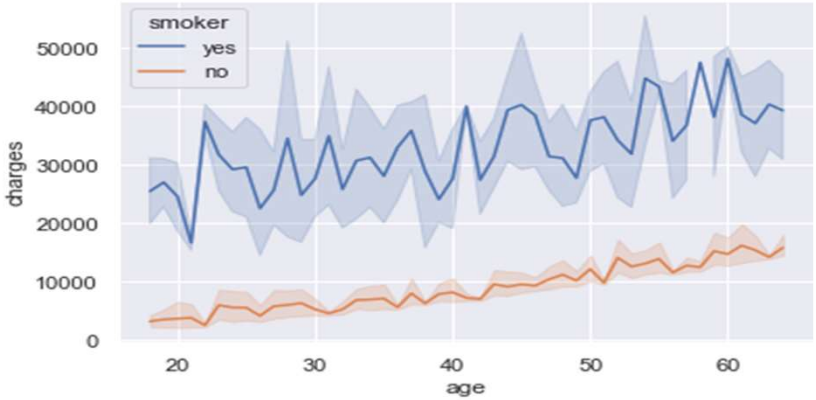
1.2 Multivariate Analysis

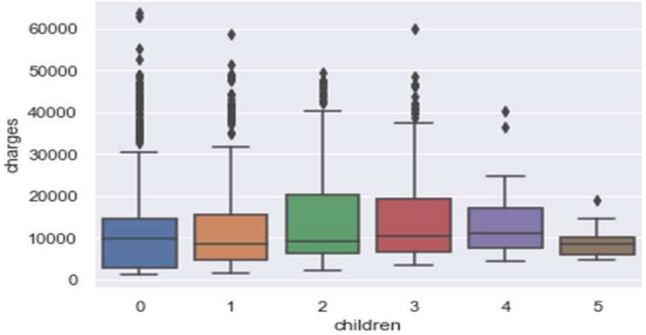












- Males have more charges (medical claim) than females.
- Smoker have much higher charges than non-smokers. For non-smokers, males have higher charges than female.
- Charges go up until 3 dependents and then falls
- As BMI increases, charges increases for smokers.
- Charges higher for smokers compared to non-smokers of the same age.
- Charges increases as the age increases.

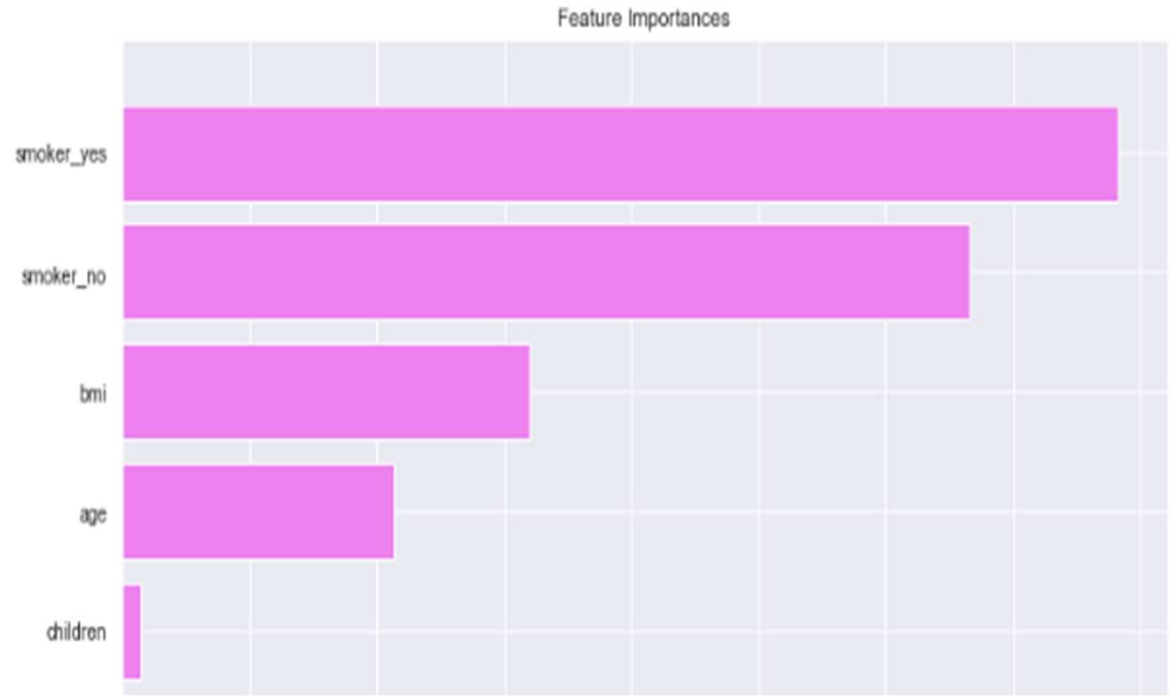
1.4 Model- Medical Claim

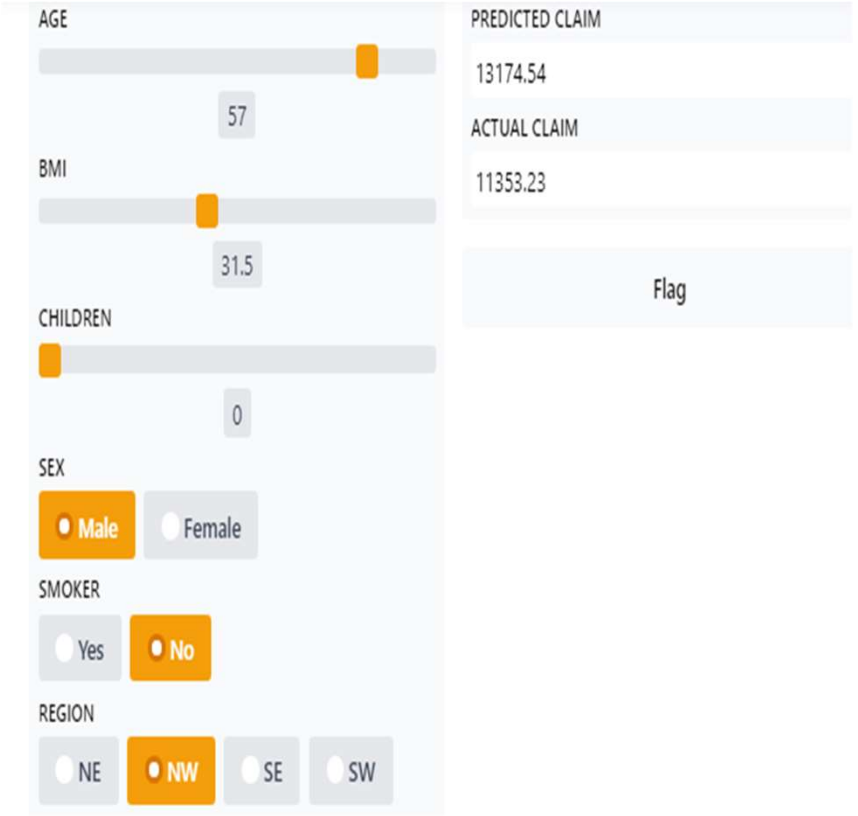
- Below is the summary of all the models created to predict the medical claim.
- Models were rejected for the following reasons:
  - ✓ models marked in amber have overfitting
  - ✓ model marked in blue had low R2 score
- Model marked in green is the best model based on low overfitting and performance.

Model	R2	
	Training Data	Test Data
Linear Regression	75.5	74.0
Decision Tree Regression- Base	100	74.0
Decision Tree Regression- Hypertuned	87.8	85.0
Random Forest Regression- Base	97.5	82.5
Random Forest Regression- Hypertuned	87.7	85.6
Gradientboost- base	90.5	85.9
Gradientboost- Hypertuned	88.9	85.9
XGBoost- base	99.6	79.9
XGBoost- Hypertuned	87.6	85.3

1.4 Model- Medical Claim

- Below chart shows the feature importance for the hypertuned random forest model used to predict the medical claim. The key factors determining the claim is whether the customer smokes, the bmi and age.
- The screenshot on the right is an interface created using gradio to utilise the above model to predict the claim based on the input data provided. This also shows the actual claims from the dataset provided.

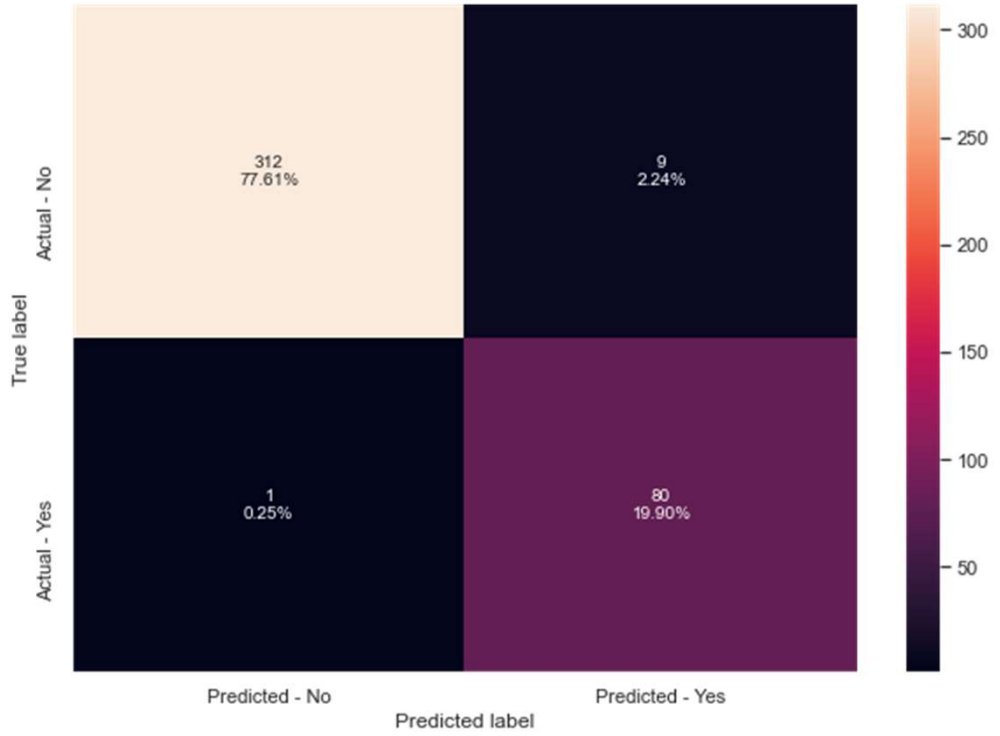




1.5 Model- Smoker

- Below is the summary of all the models created to predict whether the customer smokes.
- Models were rejected for the following reasons:
  - ✓ models marked in amber have overfitting
  - ✓ models marked in blue had low F1
- Model marked in green is the best model based on low overfitting and reasonable performance. The confusion matrix using this model is also given below.

Model	F1 Score		Recall	
	Training Data	Test Data	Training Data	Test Data
Decision Tree- Base model	100	91.9	100	91.3
Decision Tree- Max-depth=3	93.0	91.9	100	98.7
Decision Tree- Hyperparameter tuning (pre pruning)	96.4	94.1	99.4	98.7
XGBoost- - hyperparamter	100	93.5	100	88.8



1.5 Model- Smoker

- Below chart shows the feature importance for the hypertuned decision tree model used to predict whether the customer is a smoker. The key factors determining whether the customer is a smoker is the charges (medical claim), bmi and age of the customer.
- The screenshot on the right is an interface created using gradio to utilise the above model to predict the whether the customer is a smoker based on the input data provided. This also shows that model has rightly predicted that the customer is a smoker.