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

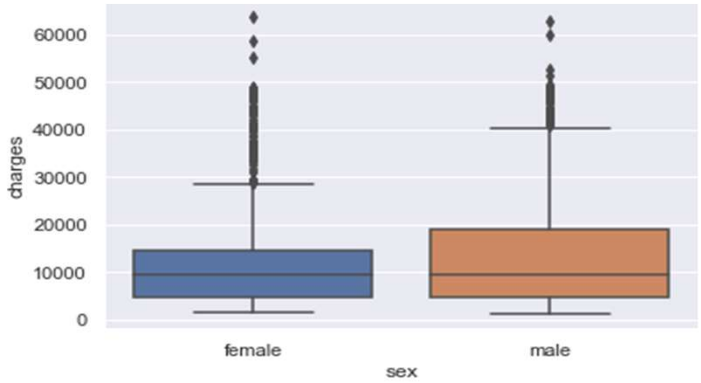
Attribute	Description	Type
Age	Age of the primary beneficiary	Integer
Sex	Policy holder's gender	Object
BMI	Body mass index (ideal BMI is within the range of 18.5 to 24.9)	Float
Children	Number of children / dependents covered by the insurance plan	Integer
Smoker	Yes or No depending on whether the insured 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

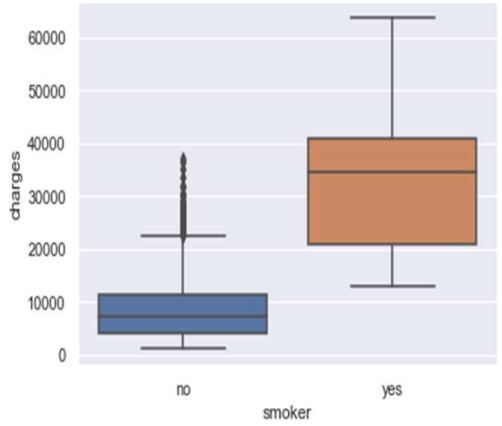
Objectives

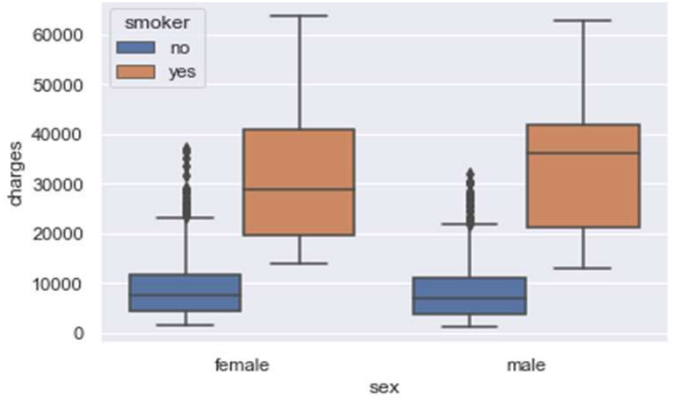
Create ML Models to:

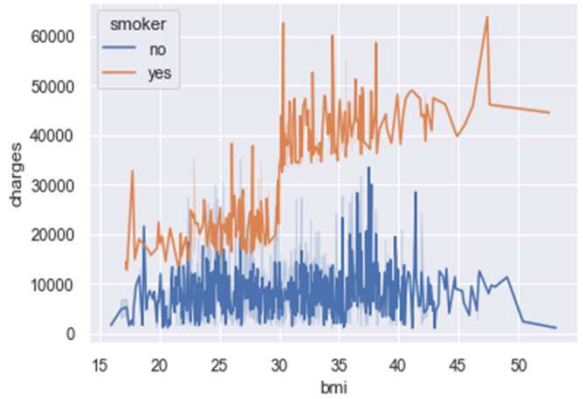
1. Predict medical claim
2. Predict whether a customer is a smoker

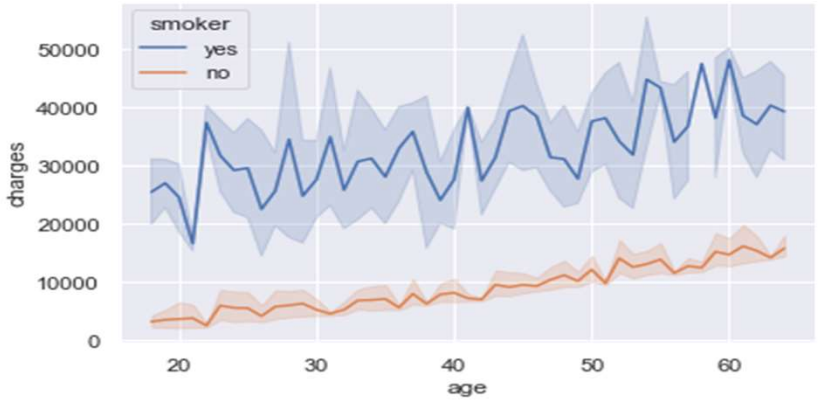
1.2 Multivariate Analysis

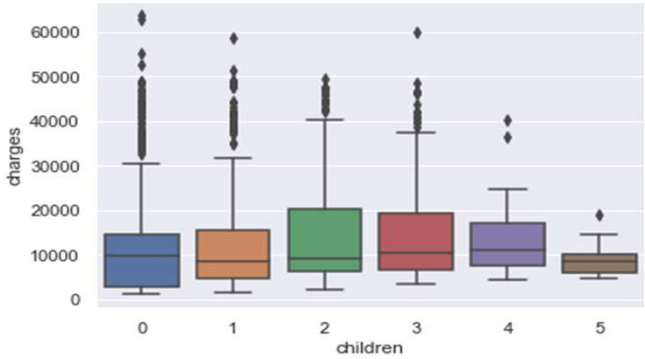












- Males have more charges (medical claim) than females.
- Smoker have much higher charges than non-smokers. For non-smokers, males have higher charges than female.
- Charges go up until 3 dependents and then falls
- As BMI increases, charges increases for smokers.
- Charges higher for smokers compared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

1.3 Model- Medical Claim

- Below is the summary of all the models created to predict the medical claim.
- Tensorflow DNN regressor gave the least overfitting.
- The tensorflow model had 4 layers of 6 nodes each. Batch size was 100 using 51k steps to train.

Model	R2	
	Training Data	Test Data
Decision Tree Regression- Hypertuned	87.8	85.0
Random Forest Regression- Base	97.5	82.5
Gradientboost- Hypertuned	88.9	85.9
XGBoost- Hypertuned	87.6	85.3
Random Forest Regression- Hypertuned	87.7	85.6
TF DNN Regressor	**85.5**	**85.7**

1.3 Loss Function

- Below is the loss plotted for different combination of batch sizes, step size and number of hidden layers for TF DNN regressor.

1.4 Model- Smoker

- Below is the summary of all the models created to predict whether the customer smokes.
- Models were rejected for the following reasons:
  - ✓ overfitting
  - ✓ low F1
- TF Linear Classification gave least overfitting

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
XGBoost- - hyperparamter	100	93.5	100	88.8
Decision Tree- Hyperparameter tuning (pre pruning)	96.4	94.1	99.4	98.7
TF LinearClassification	**96**	**96**	**96**	**96**