SC1015 Mini Project

Medical Insurance in US

Members:
Eugene Ian Teo-Aldridge (U2222253B)
Jiang Yifei(U2221092F)
Huang Shu yang (U2210979A)

Table of contents

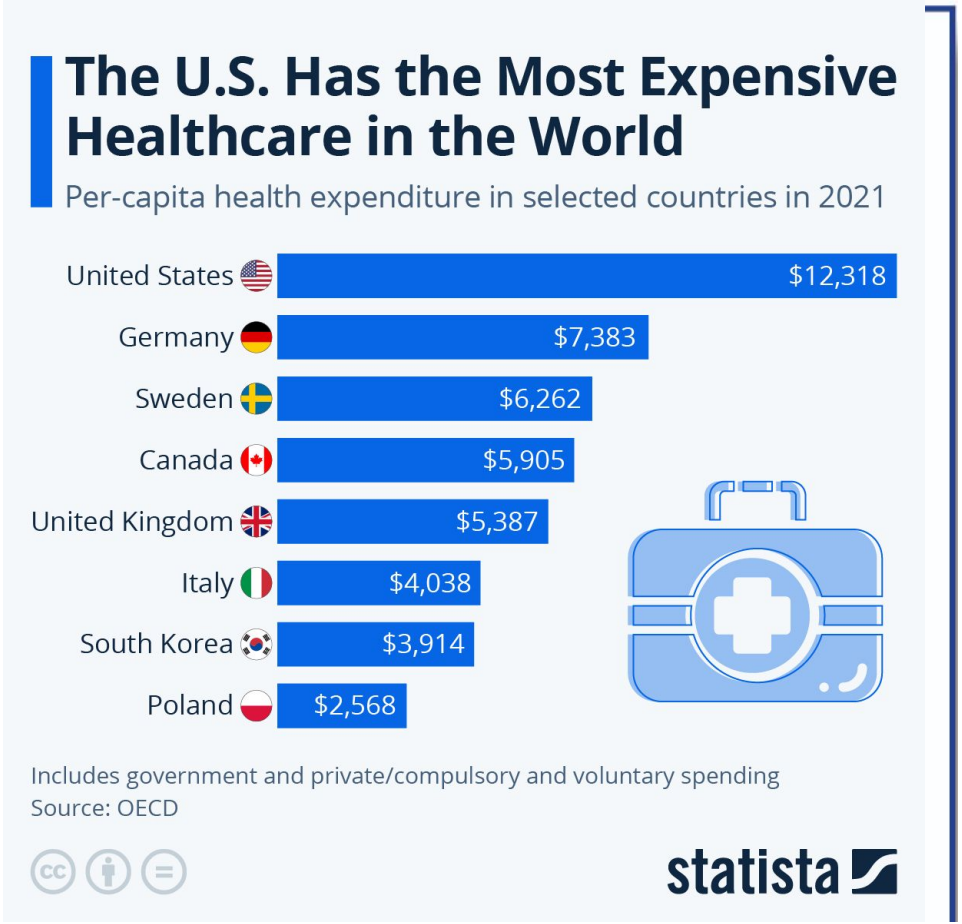
01

Problem Definition & Dataset

Expensive healthcare in US

Most common cause of personal bankruptcies in US

Medical insurance protects against high medical costs



Medical Insurance providers

Many different medical insurance plans of various premium prices

Slight mismatch in premium prices can result in loss of customers

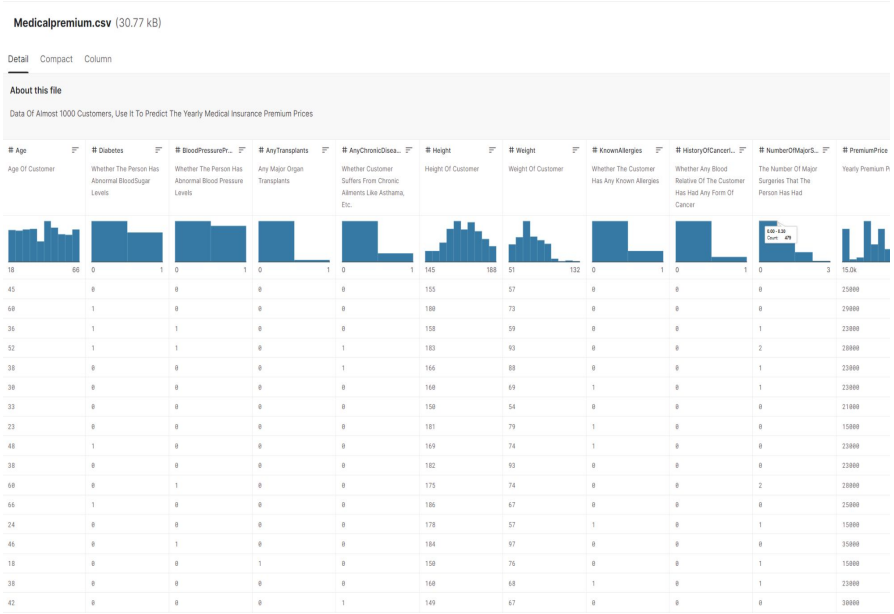
How can we develop an adaptive premium pricing model?

For medical insurance companies to remain competitive

For medical insurance policy holders to choose the best value for money plans

Dataset:
Medical Premium in the US by Tejashvi

Many numerical and categorical variables for the different health and demographic factors



Medicalpremium.csv (30.77 kB)

Detail Compact Column

About this file

Data Of Almost 1000 Customers, Use It To Predict The Yearly Medical Insurance Premium Prices

# Age	# Diabetes	# BloodPressurePr...	# AnyTransplants	# AnyChronicDisea...	# Height	# Weight	# KnownAllergies	# HistoryOfCancerI...	# NumberOfMajorS...	# PremiumPrice
Age Of Customer	Whether The Person Has Abnormal BloodSugar Levels	Whether The Person Has Abnormal Blood Pressure Levels	Any Major Organ Transplants	Whether Customer Suffers From Chronic Ailments Like Asthama, Etc.	Height Of Customer	Weight Of Customer	Whether The Customer Has Any Known Allergies	Whether Any Blood Relative Of The Customer Has Had Any Form Of Cancer	The Number Of Major Surgeries That The Person Has Had	Yearly Premium P
18 66	0 1	0 1	0 1	0 1	145 188	51 132	0 1	0 1	0 3	15.0k
45	0	0	0	0	155	57	0	0	0	25000
60	1	0	0	0	180	73	0	0	0	29000
36	1	1	0	0	158	59	0	0	1	23000
52	1	1	0	1	183	93	0	0	2	28000
38	0	0	0	1	166	88	0	0	1	23000
30	0	0	0	0	160	69	1	0	1	23000
33	0	0	0	0	150	54	0	0	0	21000
23	0	0	0	0	181	79	1	0	0	15000
48	1	0	0	0	169	74	1	0	0	23000
38	0	0	0	0	182	93	0	0	0	23000
60	0	1	0	0	175	74	0	0	2	28000
66	1	0	0	0	186	67	0	0	0	25000
24	0	0	0	0	178	57	1	0	1	15000
46	0	1	0	0	184	97	0	0	0	35000
18	0	0	1	0	150	76	0	0	1	15000
38	0	0	0	0	160	68	1	0	1	23000
42	0	0	0	1	149	67	0	0	0	30000

02

Data preparation and cleaning

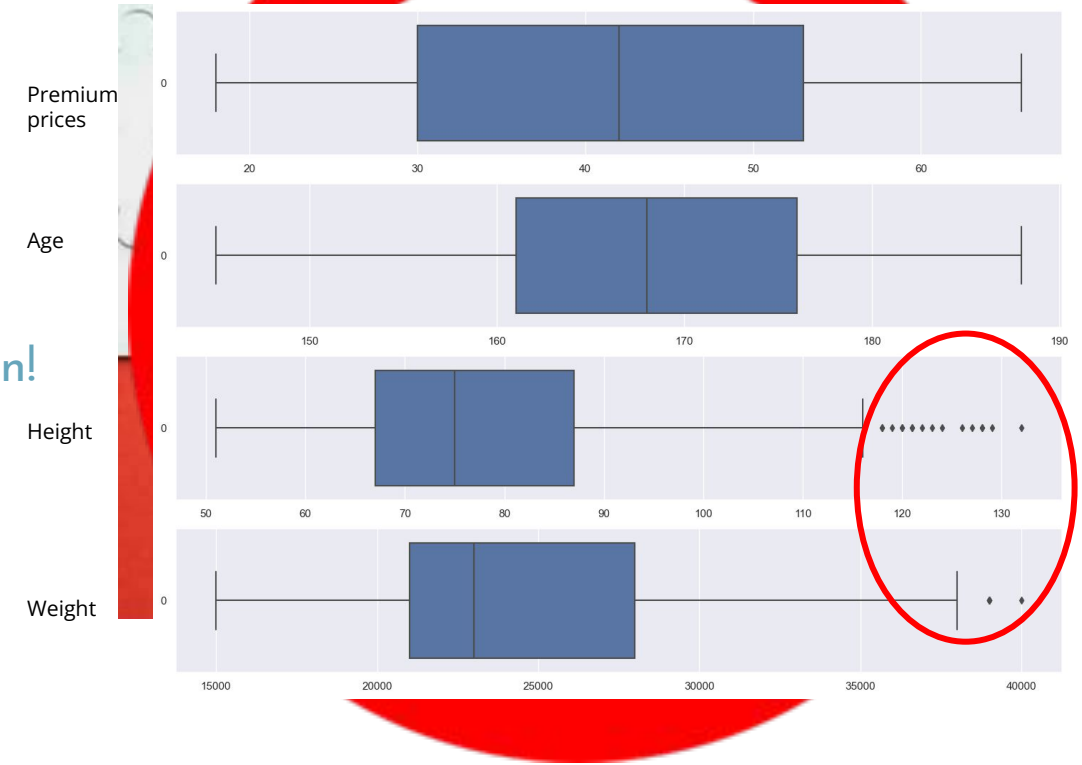
Cleaning the data

Numerical data

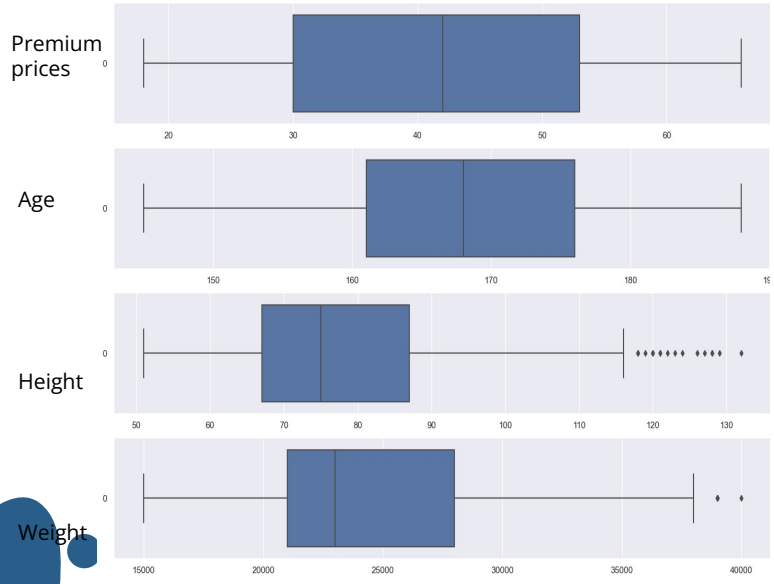
No missing data

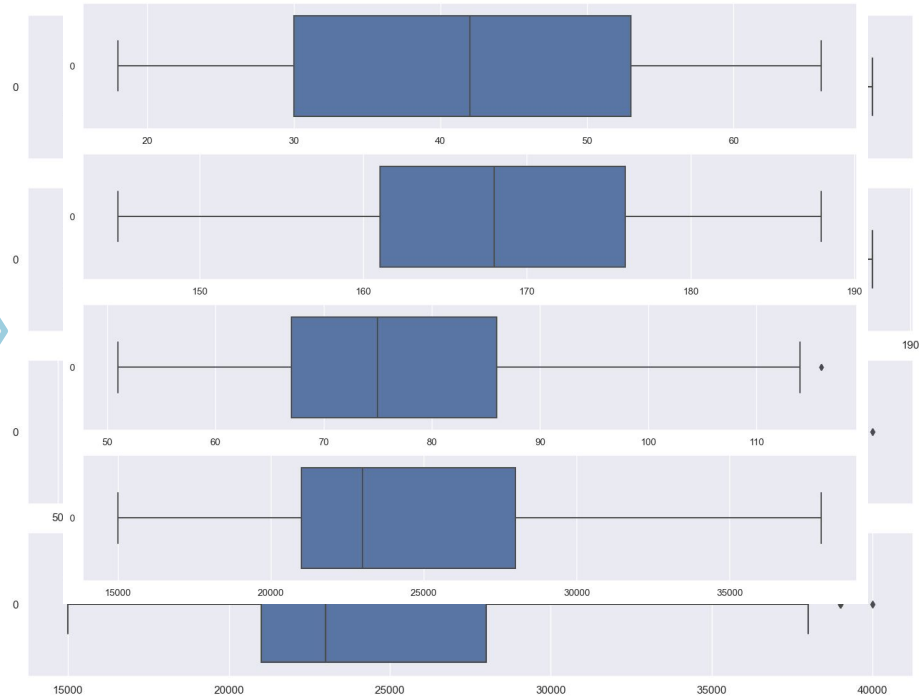
Some outliers

No need to fill in using imputation!



Removing outliers from numeric dataset



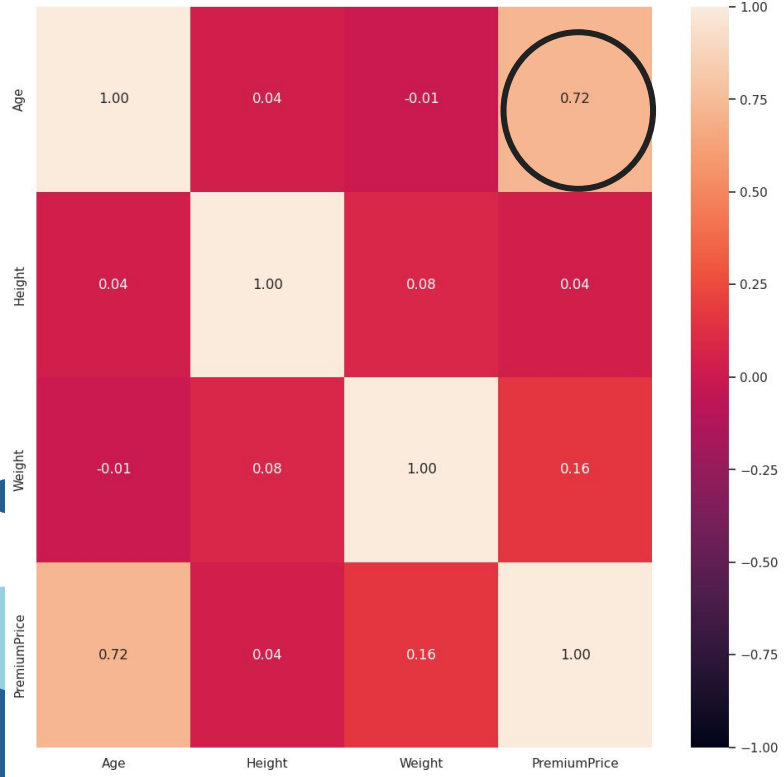


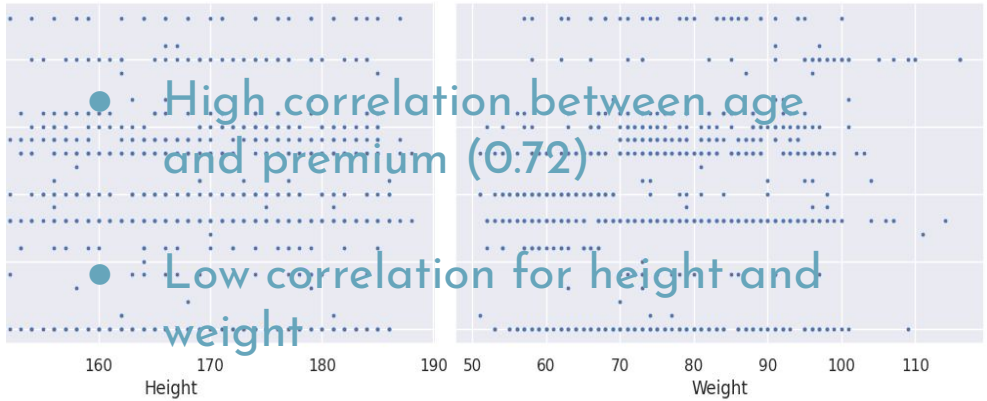
21 rows(2% of the dataset removed)

03

Exploratory Data Analysis

Comparing the numeric variables against premium price



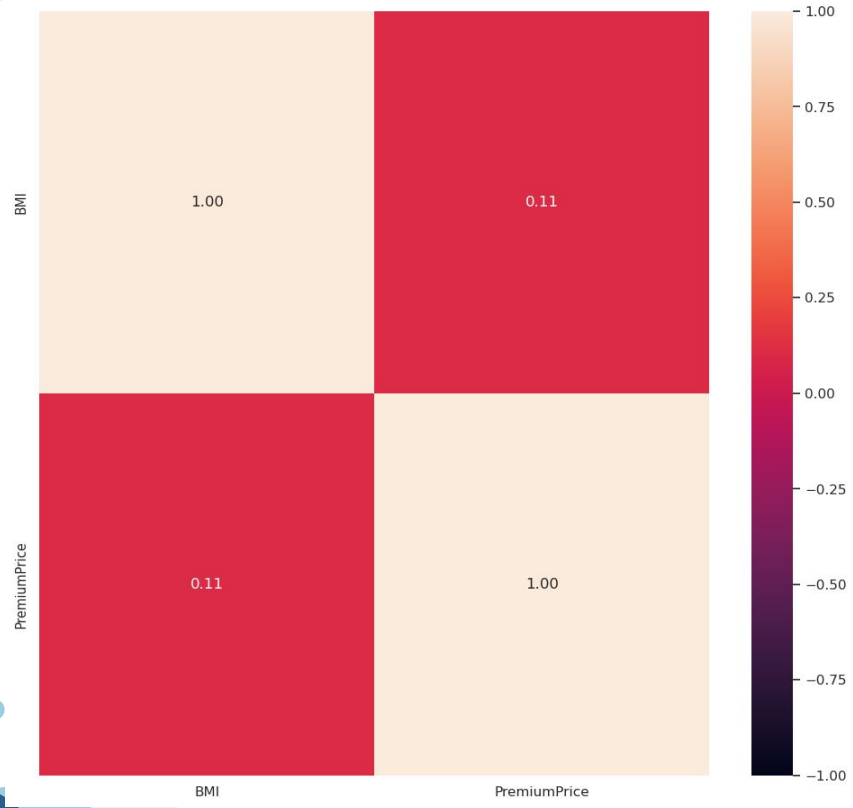


- High correlation between age and premium (0.72)
- Low correlation for height and weight

Exploring a new variable for BMI

$$BMI = \frac{\text{weight (kg)}}{\text{height (m}^2\text{)}}$$

BMI in relation to Premium prices

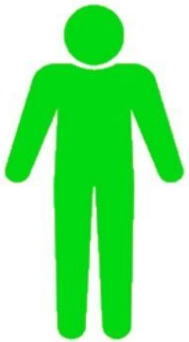


- Low correlation between BMI and premium prices (0.11)

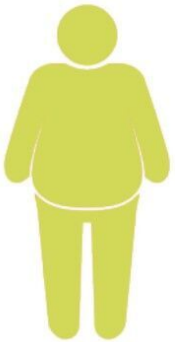
What if there is correlation different categories of BMI?

World Health Organization

Underweight:
<18.5

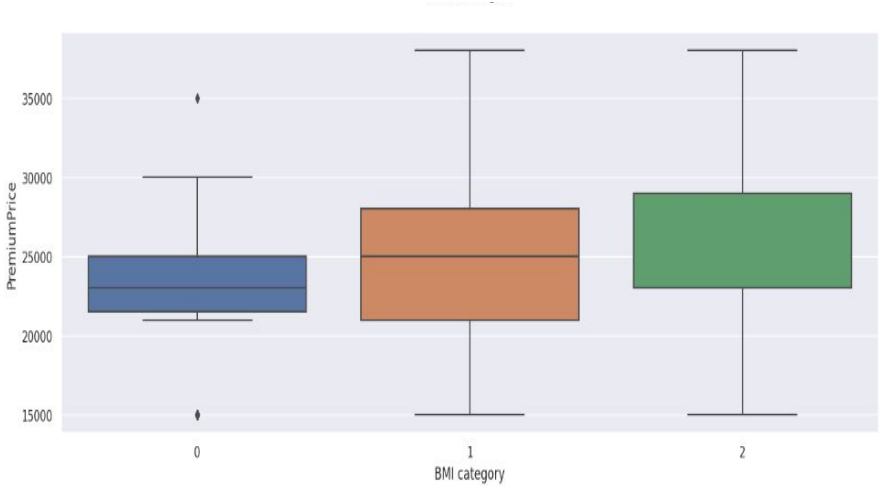


Average:
18.5 - 25



Overweight:
>=25

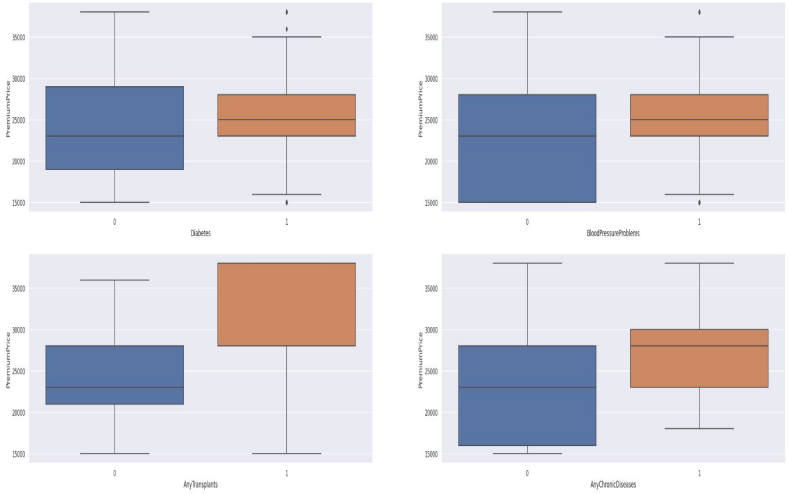
BMI Categorical findings

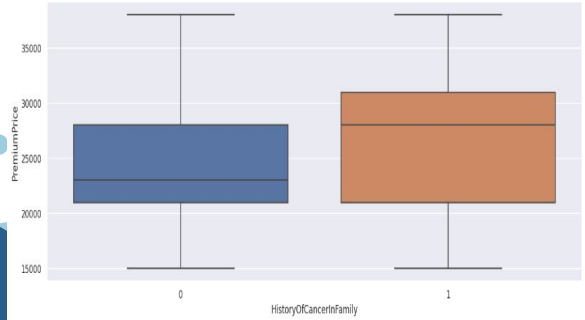


BMI Category premium price medians
Underweight: 23000
Average: 25000
Overweight: 23000

- Still low correlation between the different categories of BMI against premium prices
- Median premium prices lower for both overweight and underweight
- Not much difference in median between each categories
- This goes against our initial hypothesis that being overweight and underweight will increase premium prices

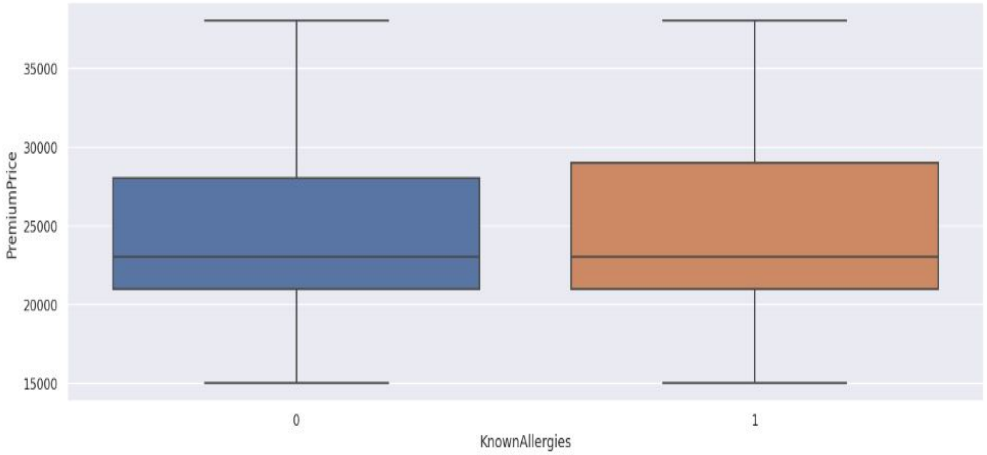
Relation between premium prices and the other categorical variables





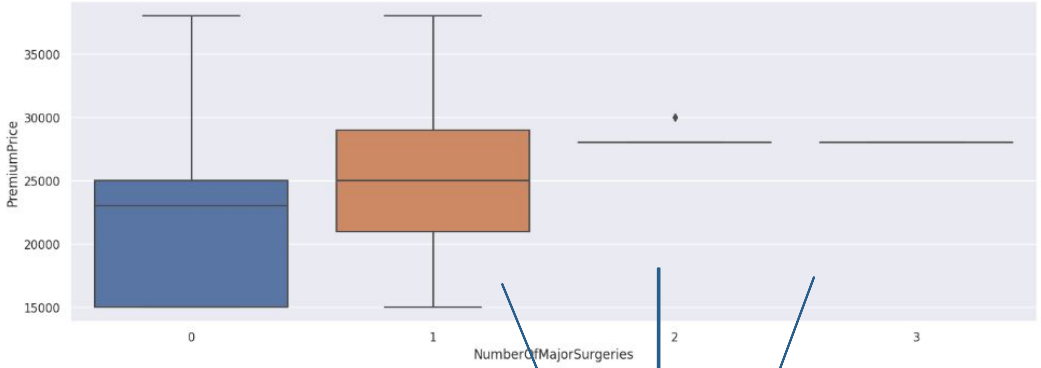
- **Diabetes**,
  Blood pressure problems,
  Any Transplants,
  Any Chronic Diseases and
  History of family cancer
  has clear relation between the different categories and premium prices

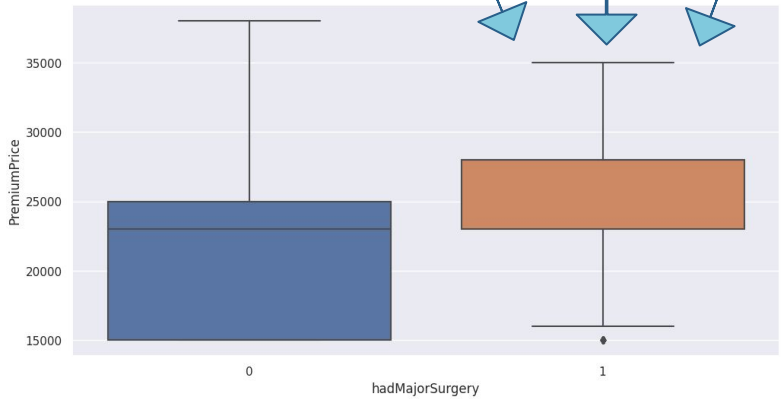
Relation between premium prices and the other categorical variables



- **Known Allergies** does not seem to have any obvious relation between its categories and premium prices

Relation between premium prices and the other categorical variables

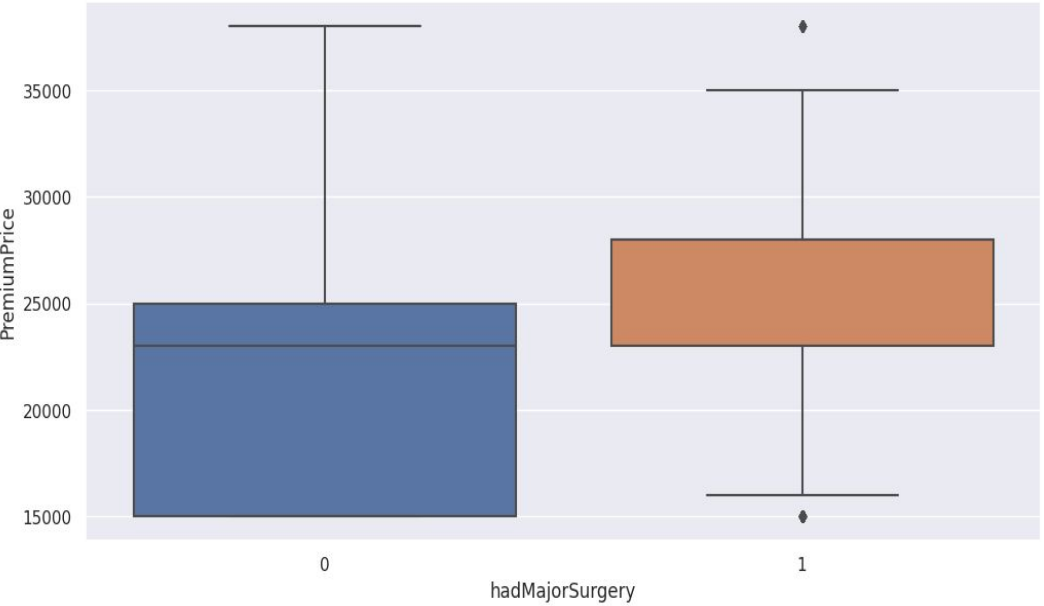




- **Number of Major Surgeries** category for 2 and 3 number of major surgeries have too little data

- Merge categories of 1, 2 and 3 together

Relation between premium prices and the other categorical variables



- **Had Major Sugery** has a clear relation between its categories and premium prices

04

Machine learning techniques

Model 1 - Random forest

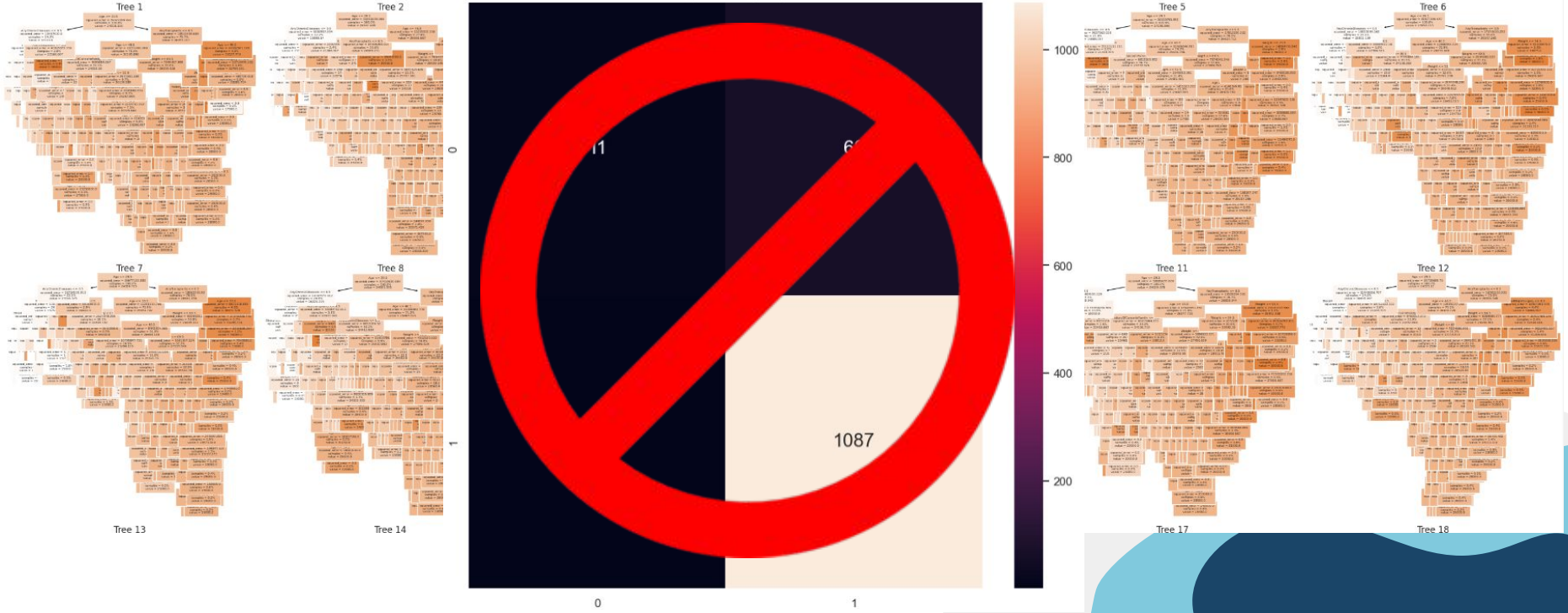
Why Random Forest:

- Random forest does not get affected as much by over fitting unlike decision trees
- Random forest is more accurate, which is important when we have so many data we are using to predict premium prices
- Has access to feature importance to measure the importance of various features

Visualisation of result

Confusion Matrix

Overfitting of Multi trees

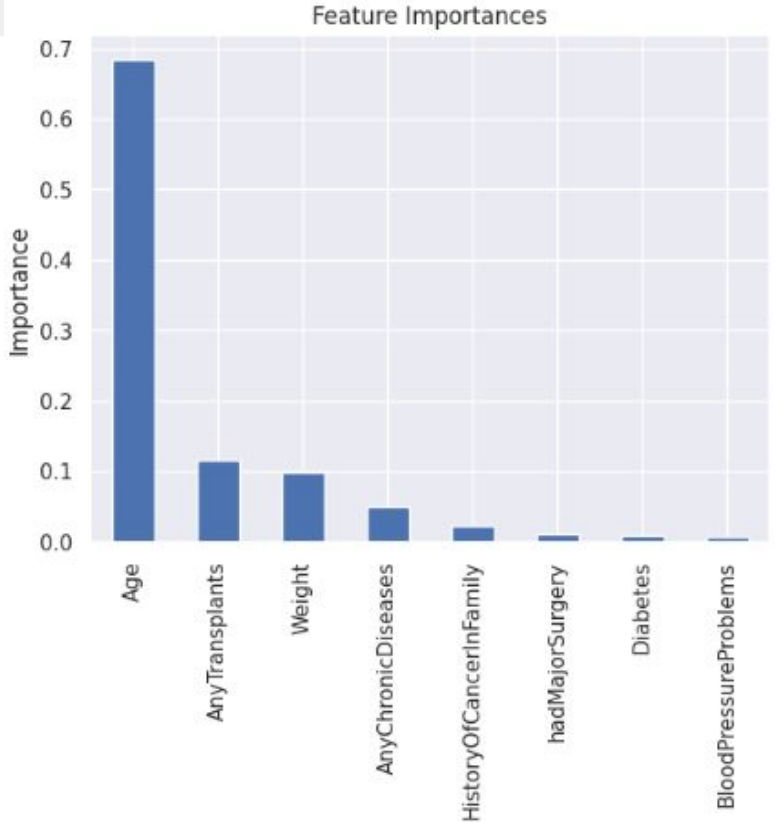


Visualisation and accuracy

Accuracy:

- R^2 value of 0.9, highly accurate predictions
- Root Mean Squared Error of 2000, fairly low given the average premium price being $24000
- Low Mean absolute error of 1000, signifying the high accuracy of the prediction

Feature Importance



Importance of features:

- Age (68%)has the highest influence in random forest price prediction
- Of the remaining values, only AnyTransplants (11%), Weight (9%) & AnyChronicDiseases (4%) influence the decision making relatively heavily
- Blood pressure issues, diabetes and having surgery barely influences pricing (<1%)

Model 2 - CatBoost

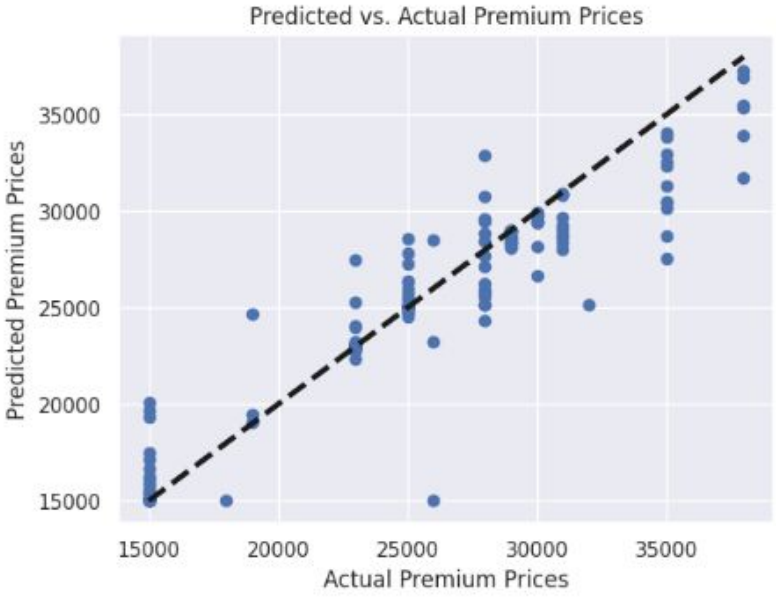
Why CatBoost:

- Specialises in prediction when provided with categorical data, which makes up the majority of our data
- Similarly to random forest, highly accurate and not susceptible to over fitting while also being resistant to noise because of the usage of gradient boosting
- Able to handle numerical data as well, while also able to provide Feature Importance

Visualisation and accuracy

Accuracy:

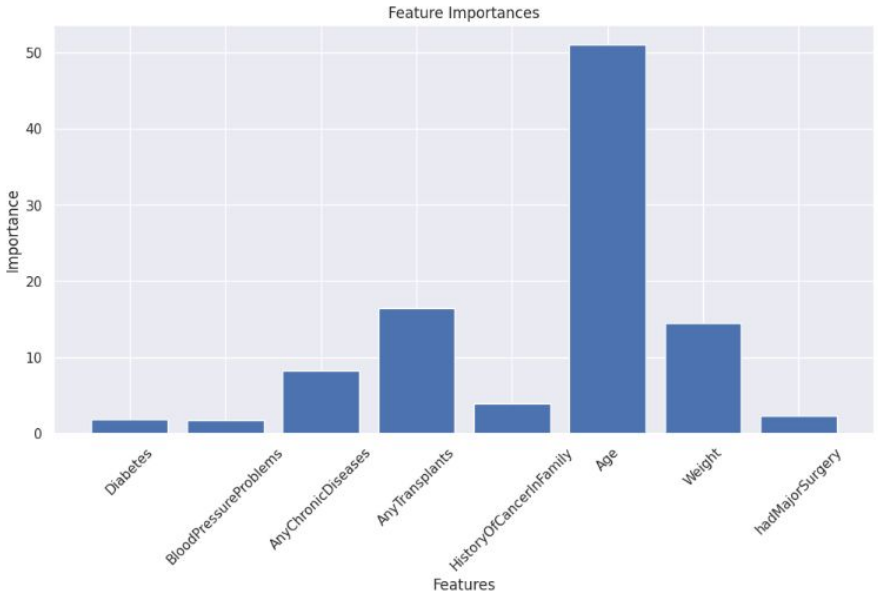
- Very Similar accuracy to Random forest, at 0.88 R^2 value
- Root Mean Squared Error of 2100 and a Mean absolute error of 1300
- Slightly less accurate than Random forest, which can potentially explained away by the nature of Catboost relying on our categorical data with lower correlation



Feature Importance

Importance of features:

- Age has lower importance this time, at 51%
- Categorical values are more important predictors, with AnyTransplants sitting at 16% and AnyChronicDiseases at 8%
- Overall, less accurate than random forest in our case, but gave more insights into the categorical values importance and correlation



05

Insights and recommendation

Lessons learnt

New Methods:

- 2 New Machine learning model, the categorical data specialised CatBoost and Random forest
- Feature engineering, making new features (BMI, BMI category) to find new things that may have strong correlations with target variable

Recommendations

What insurance companies should do:

- Understand which factors have most influence in the premium price in the insurance market and set their prices accordingly based on these features

- Remove factors with low importance like diabetes and high blood pressure to make insurance more appealing, expanding the customer base without impacting profitability

Thank you!