Medical Insurance Forecast

By: Truc Tran

Overview and Context

Context: Due to the complexities of individual medical conditions, accidents, technology, and geography, medical expenses are difficult to estimate. As a result, American health insurance firms are battling to determine how much to charge for insurance coverage. To make a profit, they must accurately forecast medical bills for their clients.

Problem Identification

Planning: In order to earn a profit, they must properly predict medical expenses for its clients. The health insurance companies decide to fund and invest to creating models, and using a collected dataset from the beneficiary's residential area in the US to predict the insurance cost for an individual.

Goal: Develop the accurate medical insurance cost forecast system for clients in the US.

Data Set

Original:

7 columns, and 1338 rows
- Independent variables (X): age, sex, bmi, children, smoker, region
- Predicted value (y) : charges

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200

Add two new features:

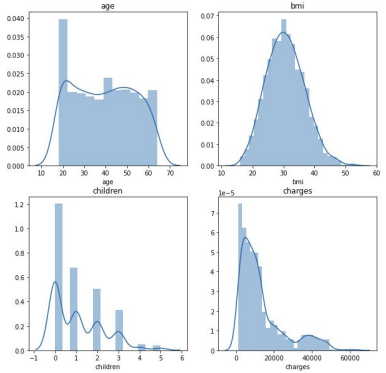
	age	sex	bmi	children	smoker	region	charges	weight_status	age_group
0	19	female	27.900	0	yes	southwest	16884.92400	2.Overweight	19-34
1	18	male	33.770	1	no	southeast	1725.55230	3.Obese	18-19
2	28	male	33.000	3	no	southeast	4449.46200	3.Obese	19-34
3	33	male	22.705	0	no	northwest	21984.47061	1.Normal	19-34

Get Dummy:

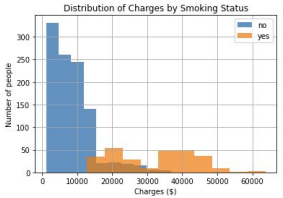
	age	bmi	charges	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	region_southwest	weight_st
0	19	27.900	16884.92400	1	0	0	1	0	0	0	1	
1	18	33.770	1725.55230	0	1	1	0	0	0	1	0	
2	28	33.000	4449.46200	0	1	1	0	0	0	1	0	
3	33	22.705	21984.47061	0	1	1	0	0	1	0	0	

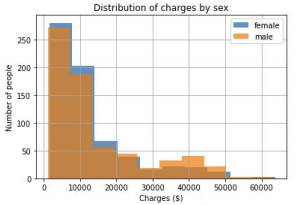
Explore Data Analysis

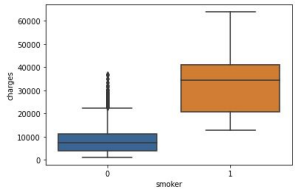
Distributions:

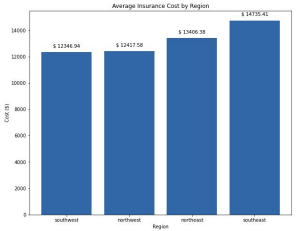


Numerical Features





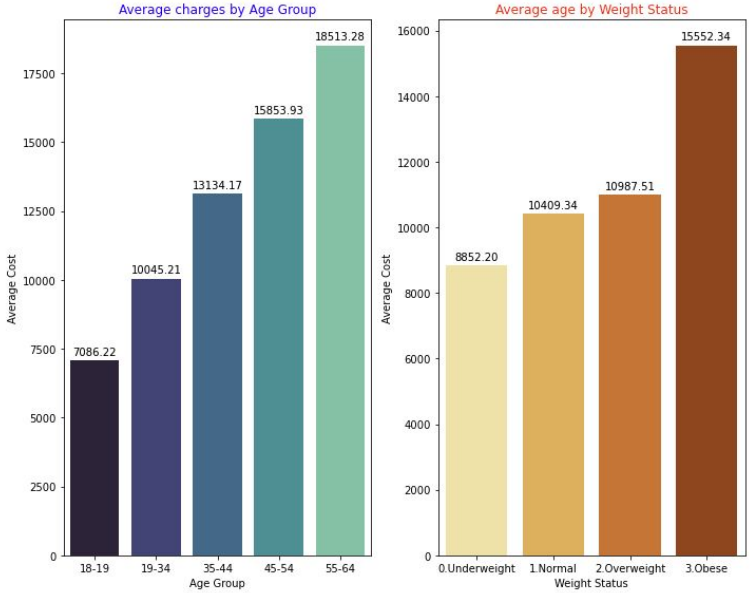




Categorical Features

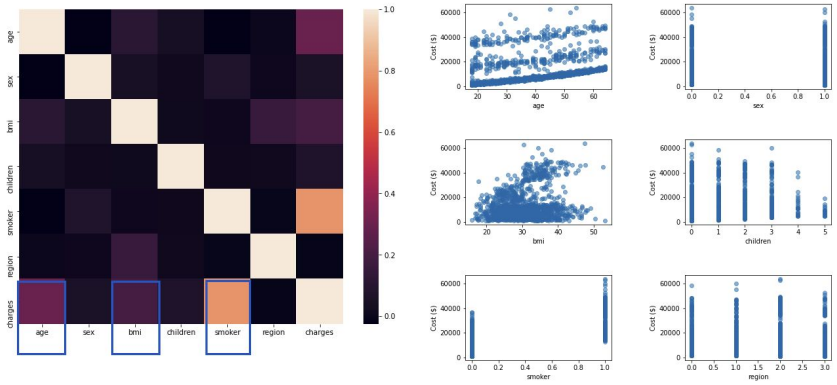
New features

Two new features: age_group, weight_status



Important factors affecting the cost

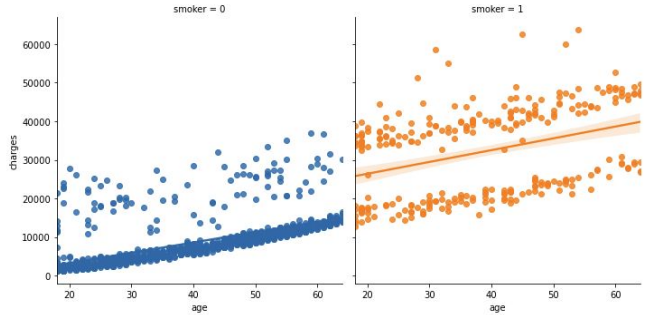
- Top 3 features: smoker, age, bmi



Heatmap

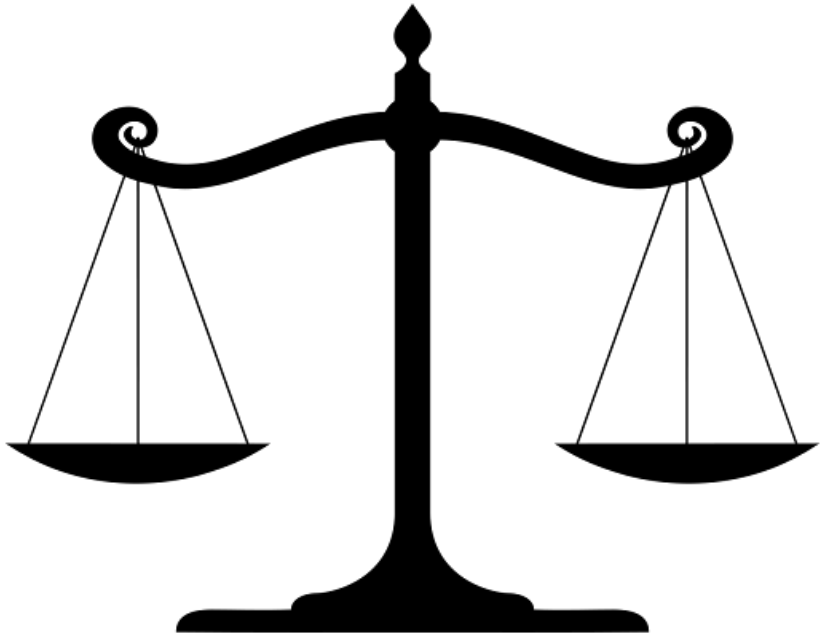
Scatter plot

The relationship between medical cost with other features



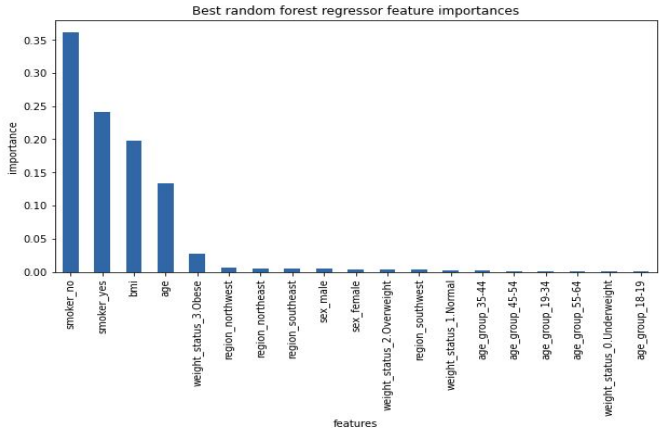
Models Comparison

	R2	Mean Absolute Error	Root Mean Squared Error
Linear Regression	0.799771	4008.447484	5536.117762
Random Forest	0.860833	2736.647824	4615.413312
Lasso Regression	0.796857	4021.451883	5576.260168



Model selection and results:

- The average cost prediction: $13,590.31
- The actual average cost: $13,100.04
- Mean absolute error: $2736.64



Recommendations

- Health insurance firms must evaluate the three major characteristics that influence cost: smoking status, age, and BMI of their clients, particularly those who smoke, are older, and are obese.
- Utilize the random forest model to forecast their clients' medical costs.

Suggestion & Future Work

- Collecting more data on medical expenses
  - Clients medical condition, disease
  - Occupation
  - Year of expenses
- Combine with another dataset from the same time period regarding the economy, GDP, and pandemics.