RUTGERS

PROJECT REPORT

REGRESSION AND TIME SERIES ANALYSIS

Vishal Reddy Mekala-213007396
Manish Maddimsetty-219007841
Dileep Kumar Pothala-218008143

INTRODUCTION

Analytics in Health Insurance: With emerging technologies driving every industry, insurance is no exception. Data Science in the insurance sector aids strategic planning and helps extract actionable insights from enormous amounts of data generated.
Analysis of customer data can be leveraged for data-driven decision making thereby increasing revenue.

PROBLEM STATEMENT

Customers often want to opt for Insurance that is robust and at the same time affordable. On the other hand, organizations want to retain maximum customers and make profits. Premium quoting for potential customers can often be a challenging task.
Predictive models built on existing customer data can be utilized to make this task a win-win situation for both- customers and companies.

OBJECTIVE

The goal of this project is to conduct exploratory data analysis to make inferences based on existing customer information. The next step is to build a model that is accurate, interpretable, and consistent to predict features based on future data.
The primary metric used to compare model accuracy is Root Mean Squared Error(RMSE).
Tasks at hand:

- Exploratory Data Analysis
- Linear Regression
- Other Predictive Modeling Techniques

DATA DESCRIPTION

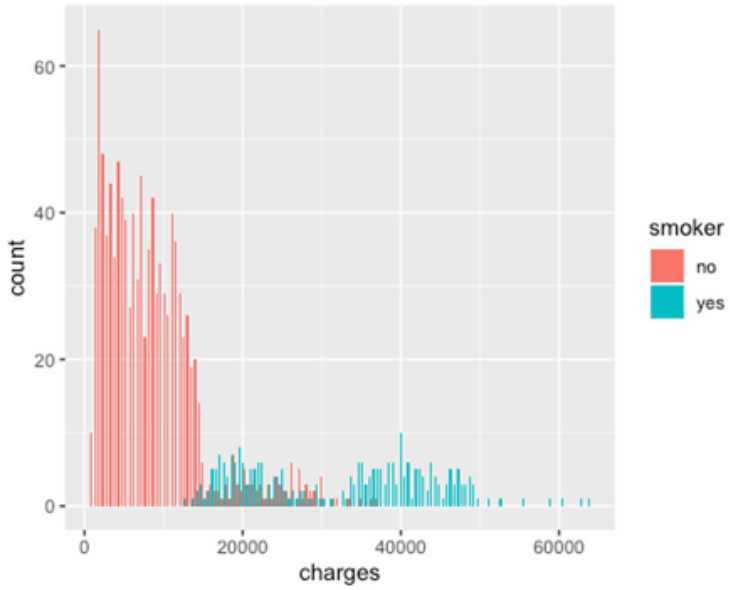
The dataset is obtained from Kaggle, and it contains information about the medical costs billed by the insurers. It contains 1338 observations and 'seven' features, and there are no missing or undefined values in the dataset. The features are as follows:

- Charges- Individual medical costs billed by health insurance(USD)
- Age- Age of an individual(Years)
- Sex- Insurance holder's gender(Male/Female)
- BMI- Body mass index of the individual(kg/m2)
- Children- Number of dependents covered by the insurer
- Smoker- If the policyholder has a smoking habit(Yes/No)
- Region- Residential area in the US(NE/NW/SE/SW)

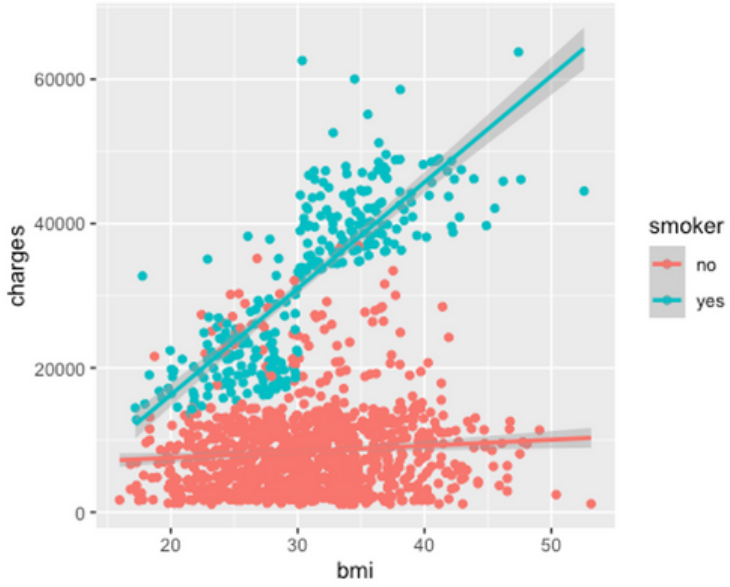
RUTGERS

EXPLORATORY DATA ANALYSIS

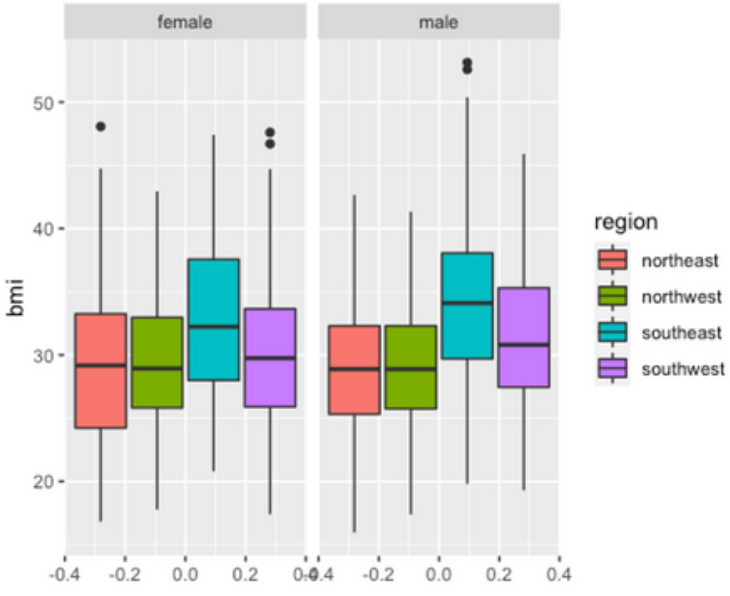
- From the below plot, it can be inferred that smokers tend to pay more charges than non-smokers. Most of the non-smokers from the data tend to pay less than $20,000.



- We can understand that for smokers, insurance charges and BMI have a linear relationship, whereas, for non-smokers, BMI doesn't seem to affect how much they are charged.
- Most non-smokers are roughly charged less than $20,000.



- People in the southeast, on average, have roughly 10% higher BMI than people in other regions, and this difference is more noticeable, especially in men.



LINEAR REGRESSION

Base Model

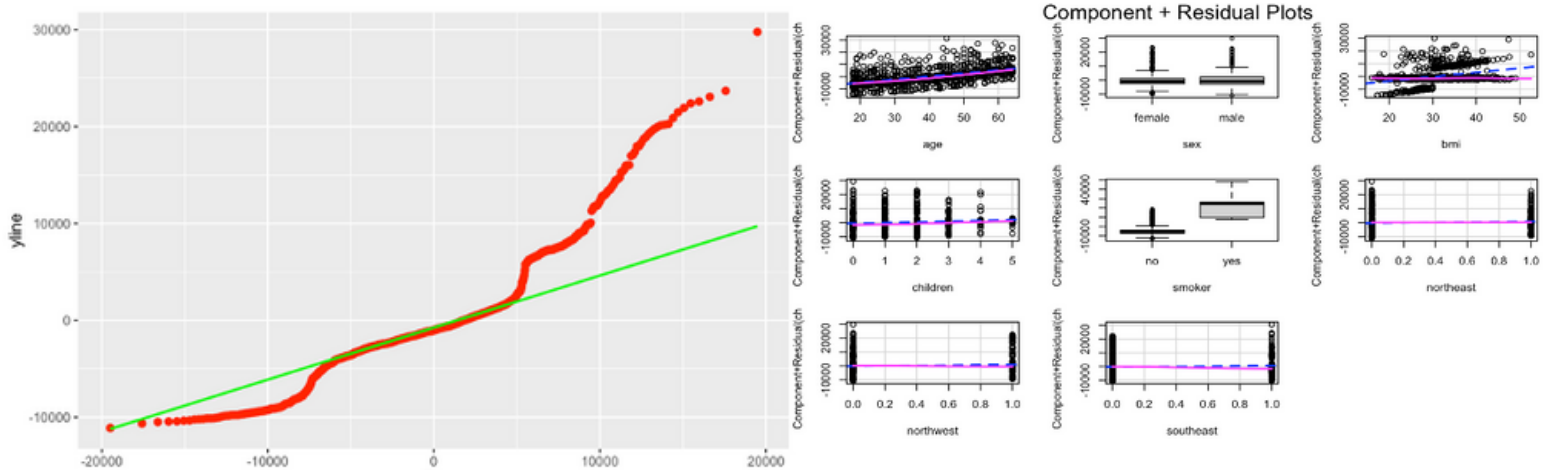
Our data has been divided into training and testing parts using the sample() function. For ease of interpretability, dummy encoding has been done on the "region" feature.

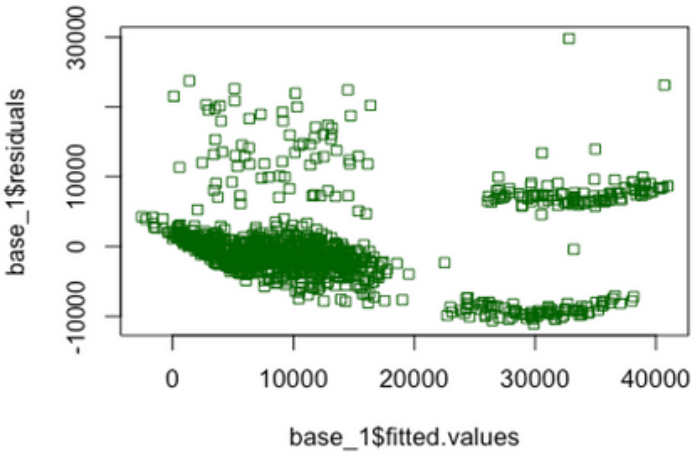
The Base Linear Regression Model has an R-squared of 75.7% and an overall statistically significant p-value.

Performing Hypothesis Testing has concluded that few of the model's coefficients are not statistically significant, but there is evidence to suggest that the overall model is significant.

Checking if the linear regression assumptions are violated has ended in proving that the linear model might not be the best fit for this data.

Normality, Linearity, and Constant Variance plots below say that the assumptions have been breached while there is evidence that Multicollinearity amongst variables is absent.





VIF SCORES:

age	sex	bmi	children	smoker	northeast	northwest	southeast
1.022053	1.011724	1.126322	1.003156	1.013885	1.529612	1.504975	1.592767

On evaluating the model on training and testing data, the RMSE is as follows:

RMSE(training)=$5919.6

RMSE(testing)=$6500

RUTGERS

Linear Regression- Cross Validation Technique

A possibility of inaccuracy in the base model could be how the data was split initially. To rule out this, the 10-fold cross-validation approach has been executed.
The RMSE is $6070.3, although not much different from the previous result; this approach ensures reliability and consistency.
The coefficients and other parameters also seem similar to what we've seen in the first model.

Linear Regression- With Interaction Terms

Categorical variables are converted to numeric columns here to make the interaction between nominal and numeric features possible. The initial model(one with many interactions) has been reduced to a simpler model. Using AIC as a criterion and performing stepwise backward selection, the best model with interactions explains about 84% of the data. On running this model on the test data, we have arrived at an RMSE= $4913.

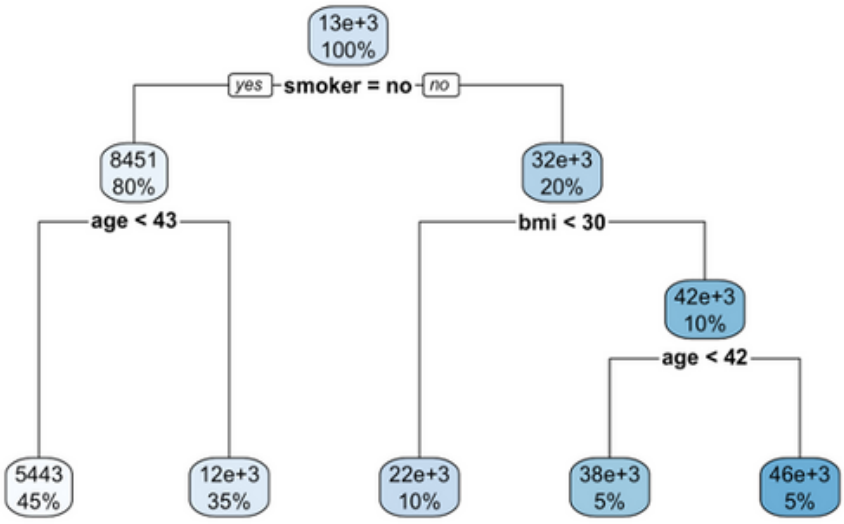
This is a significant reduction from our initial model. At this point, we have successfully built an accurate model but at the cost of interpretability. On running the diagnostic tests on this model, it has not complied with the assumptions of linear regression either.

Linear Regression-Conclusion

Based on regression diagnostics and model evaluation, it is evident that there may be better statistical models to make predictions than linear regression. We can attain a better model by possibly exploring other algorithms.

DECISION TREE REGRESSOR

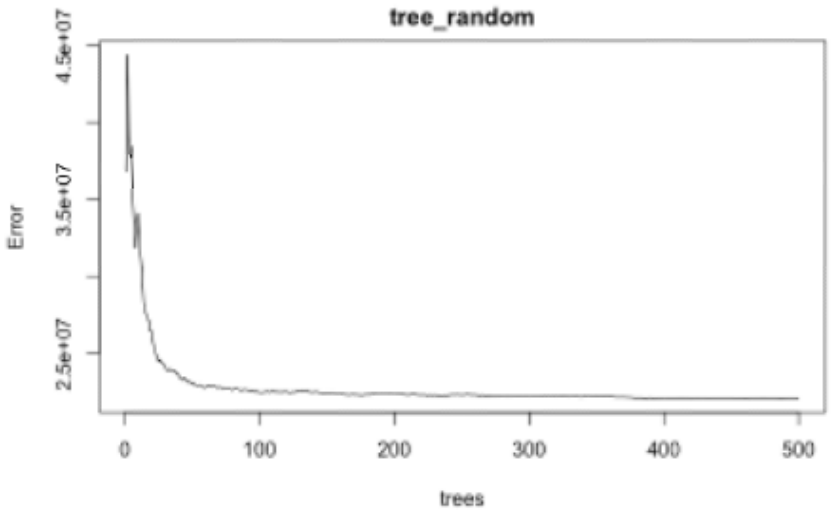
Decision trees have the advantage of interpretability but have the risk of overfitting. If the model does well on training data but fails on the testing data, then the model is said to overfit. Below is the decision tree plot for the same data.

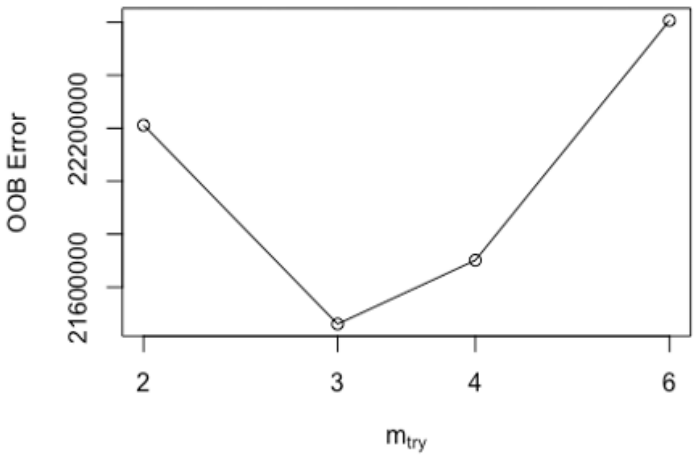


The decision tree model explains about 84% of the data and has a train RMSE of $4881.06 and test RMSE of $4912.72.
The model does perform decently on the testing data here, but it may or may not perform well on unseen data as it carries the risk of overfitting.
A more robust approach would be the Random Forest Regressor since it works on Bagging and Bootstrapping, thereby giving us consistent predictions.

RANDOM FOREST REGRESSOR

The Random Forest algorithm comes from the improvisation of decision trees. It has the advantage of interpretability, just like trees, but doesn't suffer from the risk of overfitting.
We have built two random forest models- before and after hyperparameter tuning. The first model used 'two' features at each decision node, while the final used 'three' features.
The plots below are from the hyperparameter tuning process.





The plot on the left gives us an optimal value for the number of trees used, while the right plot gives us the number of predictors to consider at a decision node.
Evaluating this model on testing data gives us an 85% explanation of the data and arrives at an RMSE of $4618.

CONCLUSION

From linear regression to random forest regressor, we have successfully increased the accuracy by 30% and, at the same time, enhanced the interpretability of the model.
Although random forest regressor stands out in terms of accuracy and consistency compared to other models in the analysis, there is more scope to this project, possibly by trying even advanced algorithms.

Link for the Data

https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

Existing Analysis

https://www.kaggle.com/code/teertha/us-health-insurance-eda

ACKNOWLEDGEMENT

We would like to thank our Professor, **Yaqing Chen,** for giving us this opportunity to work on this project. We would also like to give credit to our classmates and mentors who have guided us through this project and made it a success. It was truly a great learning experience.