Insurance Charges- a Technical report

Business Problem Statement: Earlier this year I had a major health scare and when it came time to do this project, I found Healthcare cost to be just the project. As we all know, Healthcare cost is rising rapidly and so is our Insurance charges. Having a system that will predict personalized estimated charges based on our own inputs would find a lot of customers, myself included. This project thus aims to develops a predictive model to automate premium estimation based on customer characteristics."

This report is based on the insurance dataset on Kaggle. Running the data exploration I saw that this data is clean and has no missing values. There are 1338 rows and 7 columns. As a preprocessing step, the categorical columns were encoded and created a bmi_category, and age group category. Then I performed one hot encoding for the categorical variables.

The data was then split in the 80 to 20 ratio for training and testing. The training data was featured scaled. The data was then trained on the following models: Linear Regression, Random Forest, Lasso and Ridge. Random Forest was the most consistent model on the 5 fold cross-validation with the lowest RMSE =4900, Best accuracy and smallest error bars. Then the models were trained and tested. Random Forest was the best performing model with the lowest scores for MSE(2072), RMSE(4555), MAE(2565) and the highest R2 score(0.866). Thus, the full consensus what that Random Forest was our best model by all metrics. From the Feature Analysis, it was discovered that being a smoker had the biggest impact on your insurance charges by a large margin. This is followed by bmi, so the higher rung you are in the bmi category, the higher was charges. Age also had an impact.

Thus, the key insight was that the biggest contributor to a potential higher insurance premium is your smoking status. A smoker is predicted to be a high risk. Of course, the older you get the higher is the insurance charges prediction as well. Finally, one other factor that also has more affect on a higher premium was the gender. A male is placed on higher risk than a female.

The biggest limitation for this project is that the data is not recent. So although the mechanics of the app is working, the charges predicted is not in tune with the times. The biggest improvement would be having access to a recent data source . That would improve the project's overall effectiveness.

In conclusion, this Random Forest model achieves 86.6% prediction accuracy (R2), significantly outperforming traditional regression methods. This translates to predictions that are closer to actual customer value on average. This will enable more precise business decision making.