William Barker
DSC630
Final Project

Predicting Health Insurance Charges

Introduction

The United States healthcare system is a controversial topic for many people. Much of our healthcare is privatized and expensive medical costs can sink a family financially if they don’t have decent health insurance. This makes owning health insurance essential for Americans who want to ensure that a trip to the hospital doesn’t bankrupt them, but not everyone in the country is treated equally when it comes to how much they have to pay for certain health insurance plans. This project aims to look at costs from the insurance companies perspective, and I will try to see if I can use data to figure out how much money an insurance holder will cost the company, therefore determining if they should be charged more or less for insurance.

Data Selection and Preparation

For this project I will be using a dataset from Kaggle called “Medical Insurance Cost Prediction.” The dataset has 2,700 rows and seven columns, the columns being:

1. Age

2. Sex

3. BMI (Body Mass Index)

4. Children

5. Smoker

6. Region

7. Charges

“Age” is the policy holders age, “Sex” is their sex (male or female), “BMI” is their body mass index, “Children” is how many children they have, “Smoker” is if they are a smoker or not, “Region” is what region of the United States they live in, and “Charges” is their monetary medical charges. For the data preparation part of my project I didn’t have to do too much. I dropped any NaN values in the dataset and converted any non-numeric values to numbers. This was done with the “sex”, “smoker” and “region” columns by mapping their string values to either 1, 2, or 3. In the sex column, male became 1 and female became 2. In the smoker column, no became 1 and yes became 2. In the region column, southwest became 1, southeast became 2 and northwest became 3.

Building Models for Analysis

For this project I built and evaluated three different models to see which would work best with my data and the goal of my project, which is to try and predict a person's health insurance charges. I chose to run a linear regression model, a random forest model and a K-means algorithm on my dataset. I started with the simplest model, a linear regression. To analyze my model, I checked its Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and its R-squared score (R2). I got an MAE of 3950.60, which means my linear regression prediction was off by $3,950.60 when compared to the actual values. The values in the charges column can vary pretty wildly, but the average value is $13,261.38, so while our error isn't terrible, it's also not super great. For the RMSE I got a value of 6154.54, which means difference between my predictions and the actual values was about $6,154.54, which is worse than the MAE score. Finally I checked the R2 score and got approximately 0.765, which means about 76.5% of the variability in charges can be explained by the features (age, sex, bmi, children, smoker, region) used in my linear regression model. This suggests that my model captures a significant portion of the variability in charges.

Next I tested a Random Forest model, looking at MAE, Mean Squared Error (MSE), RMSE and R2 Score for analysis. The results of my analysis were:

MAE: 1,256.58

MSE: 7,243,418.85

RMSE: 2,691.36

R2 Score: 0.955

These results are much better than the linear regression the results and actually rather promising! The MAE shows the random forest model was only about $1,256.58 off from the actual charges, RMSE was only $2,691.36, much lower than with the linear regression and the R2 Score indicates that this model explains about 95.5% of the variance in the target variable, which suggests that the model fits the data well and has strong predictive power!

Finally I trained a K-Nearest Neighbors regression model, and tested the same metrics. My results were:

MAE: 5,825.06

MSE: 94123020.0

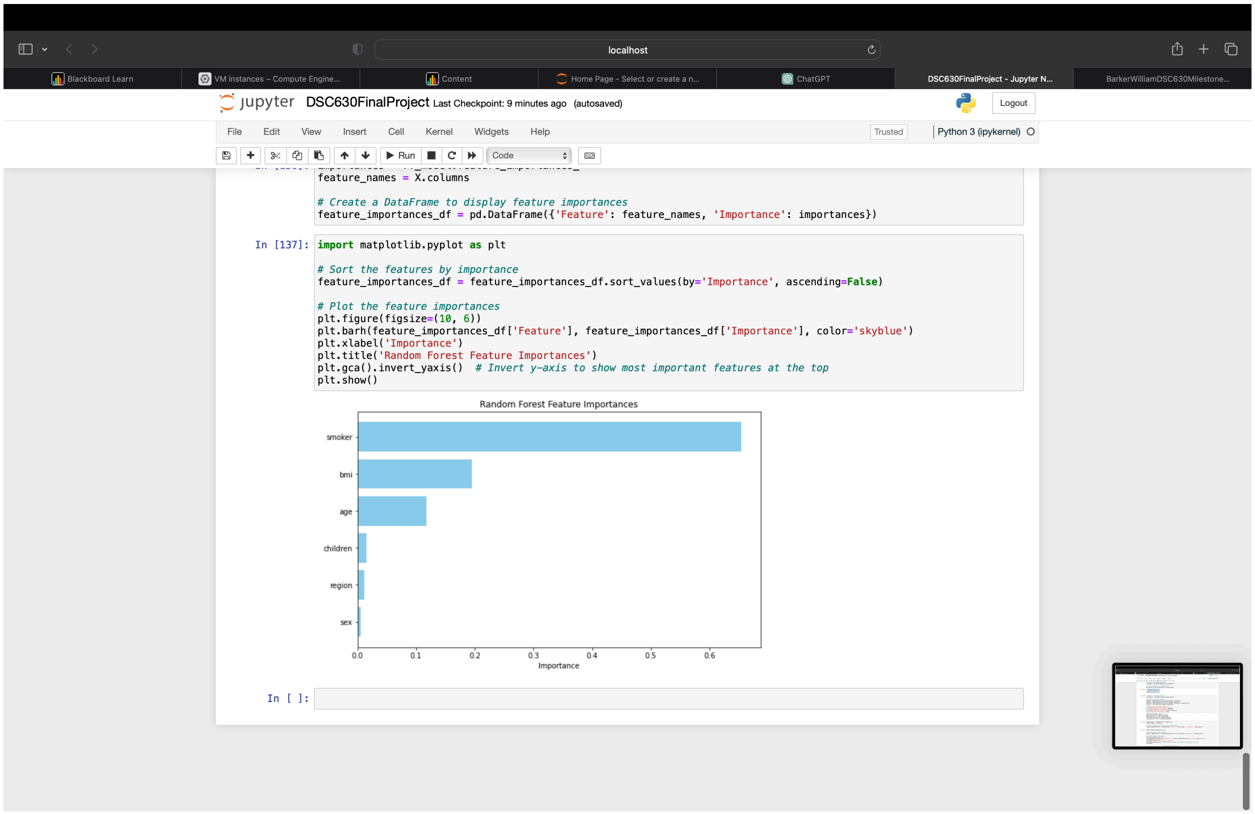
RMSE: 9701.7

R2 Score: 0.42

This model gave me the worst results so far. An R2 Score of 42% isn't terrible but is much lower than what we got from the random forest model.

Conclusion/Interpretation

The goal of my project was to try and accurately predict a persons health insurance charges using different metrics like age, BMI, sex etc. Out of the three models I tested, my random forest model was the most promising, with a 95.5% R2 Score! I also wanted to see what features of my dataset had the biggest impact on the charges column, using the random forest model as my reference. The visualization I made shows that the biggest impact on a customers insurance charges is if they're a smoker or not, followed by their BMI and their age.


localhost
Blackboard Learn
VM instances – Compute Engine...
Content
Home Page - Select or create a n...
ChatGPT
DSC630FinalProject - Jupyter N...
BarkerWilliamDSC630Milestone...
jupyter
DSC630FinalProject Last Checkpoint: 9 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code
Random Forest Feature Importances
smoker
bmi
age
children
region
sex
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance

```
feature_names = X.columns

# Create a DataFrame to display feature importances
feature_importances_df = pd.DataFrame({'Feature': feature_names, 'Importance': importances})
```

```
In [137]: import matplotlib.pyplot as plt

          # Sort the features by importance
          feature_importances_df = feature_importances_df.sort_values(by='Importance', ascending=False)

          # Plot the feature importances
          plt.figure(figsize=(10, 6))
          plt.barh(feature_importances_df['Feature'], feature_importances_df['Importance'], color='skyblue')
          plt.xlabel('Importance')
          plt.title('Random Forest Feature Importances')
          plt.gca().invert_yaxis()  # Invert y-axis to show most important features at the top
          plt.show()
```

```
In [ ]:
```

Whether a person has children and what region they live in had small impact on charges, and a persons sex had almost no impact. This result is mostly what I was expecting. Smoking has already proven to be a killer and something that destroys a persons lungs, so it's no surprise being one would lead to higher medical expenses compared to not being one. BMI is body mass index, and it's also relatively common knowledge that being over (or even under) weight can detrimental to a persons health. Age also isn't surprising, as a persons health naturally will decline in their later years. Overall I would say that this project was a successful one. Using the random forest model we can accurately predict a persons medical insurance charges with a low margin of error, and was able to pinpoint which factors played the most significant roles in making those charges. There are some ethical considerations to be made in regards to implementing data projects like this to try and determine a consumers health insurance prices. Even though a persons BMI was the second leading cause in higher medical charges, many would argue that it would be unethical to charge people with a worse BMI more than those in a healthier range, but it may be more acceptable to charge smokers more than non-smokers since smoking is something done recreationally and can be stopped with enough willpower.

References

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