Optimizing Healthcare Expenditure through Machine Learning

Michael Montana

March 2, 2024

INTRODUCTION

This report focuses on applying machine learning to the personal medical cost dataset retrieved from Kaggle.com[i]. The data set contains seven features. These features are age, sex, BMI, children, smoker, region, and charges. Age, sex, and BMI represent the biological age, sex, and body mass index of the primary beneficiary. The children feature provides a count for the number of children covered in the health care plan. Smoker provides a Boolean yes or no for the primary beneficiary's smoking status. Instances in the region feature identify one of four United States regions, northeast, southeast, southwest, and northwest. The last feature, "charges," provides individual health insurance cost. The objective for this project was to explore, analyze, and preprocess the data and apply the regression analysis to the processed dataset to the data set to predict the cost of health insurance.

BUSINESS PROBLEM/HYPOTHESIS

My hypothesis is that it may be possible to predict healthcare costs with a meaningful degree of accuracy using machine learning. I also believe that certain factors will show a positive correlation with the price of healthcare. This information can be used to inform insurance customers of their increased risk of healthcare costs, allowing them to make life changes while, at the same time, the insurance companies can adjust their rates appropriately to cover the increase in policy costs.

METHODS/ANALYSIS

This project will involve data processing, exploratory data analysis and visualization, and the building of various machine-learning models with hyperparameter tuning. The first step in this project will be data processing.

During the data processing portion of this project each data feature will be to determine its data type. Each feature will be reviewed for null values, and if any are identified, the null will be replaced with the median feature value. All categorical data, non-numeric data, will be transformed into ordinal or binary features. Categorical features being transformed ordinally will include instances with a natural ordering, such as small, medium, and large. Categorical data that is transformed into a binary feature will be done

so using One-Hot Encoding, creating new features for each unique instance of data in the former feature, and every instance in each newly created feature will be assigned a True/False (1/0) binary value. After the data is cleaned, the data analysis can begin.

During the project's exploratory analysis, numerical data of summary statistics will be reviewed, while the count and percentage of categorical data will be analyzed using the original unprocessed data. Both the summary and count statistics will be visualized to gain quick insights into the data features. Using the processed data, the features will be correlated and displayed on a heat map to show positive and negative relationships between the features. In the last phase of exploratory analysis, the features will be compared using a pair plot to visualize their relationships. If any data issues are discovered, the data will receive additional cleaning/processing, but if no data issues are discovered, the data will move forward to modeling.

The first thing that must be understood when building machine learning models is the target of the data set. If the target is numerical, a regression model is used while if the prediction target is categorical a classifier model is used. Once the target is identified and understood, the data will be split into X and y train and test data sets. The X value represents the independent variables, while the y value represents the dependent target variable. The purpose of the train and test data sets is to train the model and perform an independent test using the set-aside test data to determine the model's accuracy. The next step in the process is to select various classifier and regression model types. Each model will have various hyperparameters selected and adjusted during the fitting process to identify the highest-performing setting for each model. Also, during the fitting process, the X/y train and test datasets will be sliced in five different ways using the k-folds method to prevent data sampling bias. Once this is complete, the best-performing model will be selected. The next section of this report will contain the results from each.

RESULTS

During preprocessing, age, BMI, children, and charges were identified as numerical features and smoker and region were identified as categorical features. No null values were identified. The smoker and region

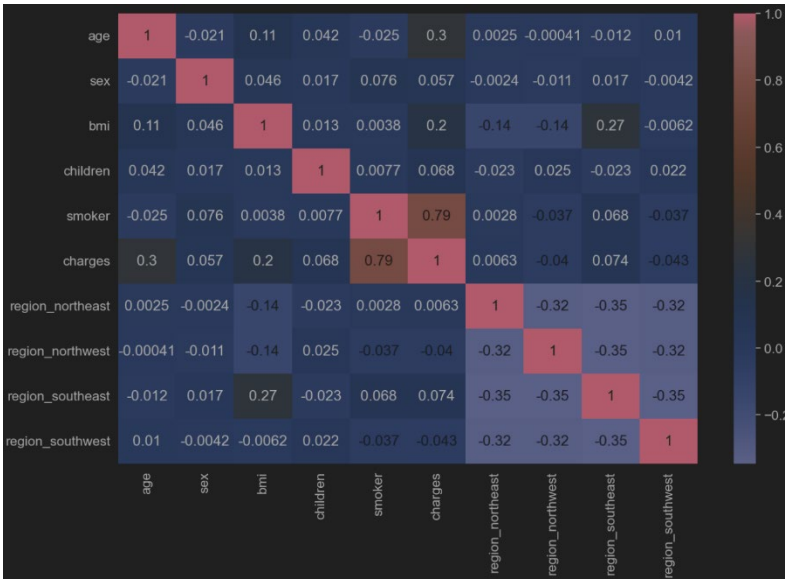
features were transformed using One-Hot encoding. Smoker remained as a single feature with a 1 or 0 for each incident of yes/no. The region feature was split into four new features, one for each of the four unique instances. The following image is a snapshot of the first five instances of the dataset.

	age	sex	bmi	children	smoker	charges	region_northeast	region_northwest	region_southeast	region_southwest
0	19	0	27.900	0	1	16884.92400	False	False	False	True
1	18	1	33.770	1	0	1725.55230	False	False	True	False
2	28	1	33.000	3	0	4449.46200	False	False	True	False
3	33	1	22.705	0	0	21984.47061	False	True	False	False
4	32	1	28.880	0	0	3866.85520	False	True	False	False

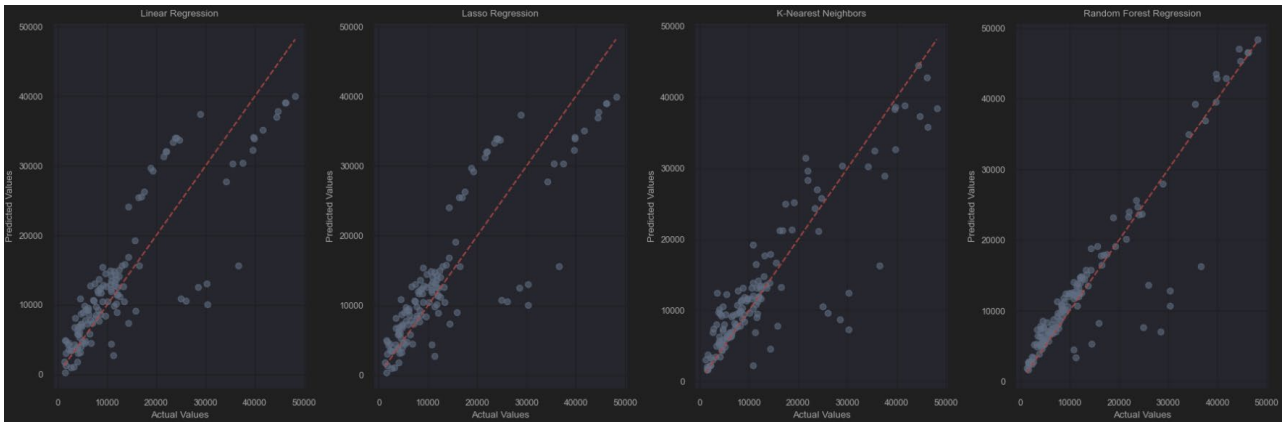
Most of the categorical features were evenly distributed. The one exception was Smoker with only 20.5% identifying as smokers and the rest identifying as non-smokers. Below is the statistical summary of the numerical features in the data set.

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

While exploring the data, the most significant finding was the correlation between independent features and the target feature (charges). As many would expect, BMI (0.2), Age (0.3), and Smoking (0.79) were the three factors with any significant impact on the charges as the figure below highlights.



Since the target feature is numerical, regression algorithms were used for the modeling process. Four model types were used: Linear Regression, Lasso Regression, K-Nearest Neighbors, and Random Forrest Regression. R-squared Score was used to compare the performance of the models. The Linear and Lasso regression models performed almost identically with R2 scores of 0.7325 and 0.7326. The K-Nearest Neighbors model performed better with an R2 Score of 0.7736. The best performing model was the Random Forrest regression model with achieved an R2 Score of 0.8437. The following image compares the predicted values against the actual test values.



CONCLUSION

The results demonstrate the ability to predict healthcare costs with a meaningful degree of accuracy using machine learning. Positive correlations between BMI, age, and smoking and higher healthcare costs were identified. This information can be used to inform insurance customers of their increased risk of healthcare costs. Insurance companies can use machine learning to better set the rates of their coverage plans.

REFERENCES

[i] Medical Cost Personal Datasets (kaggle.com)