Final Project

Modeling Insurance Data

Hunter Schuler, Li Yuan
December 6th, 2022

Overview

The data we will be analyzing is US health insurance data, publicly available on Kaggle.com. The data is licensed under a CC0: Public Domain license. The data was first posted approximately 3 years ago and comes with no documentation. The data consists of 7 features and 1338 data points. We will use 6 of these (explanatory) features in creating models to predict the 7th (response) feature, "charges."

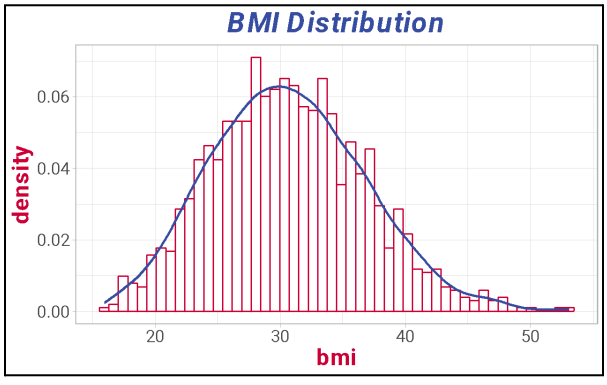
The first feature is age, an integer ranging from 18 to 64. The second feature is sex, a string with values "male" and "female." The third feature is bmi (body mass index), a number ranging from 15.96 to 53.13. The fourth feature is children, an integer ranging from 0 to 5. The fifth feature is smoker, a string with values "yes" and "no." The sixth feature is region, a string with values "southeast," "southwest," "northwest," and "northeast." The last feature (the response variable) is charges, a number ranging from 1121.874 to 63770.43.

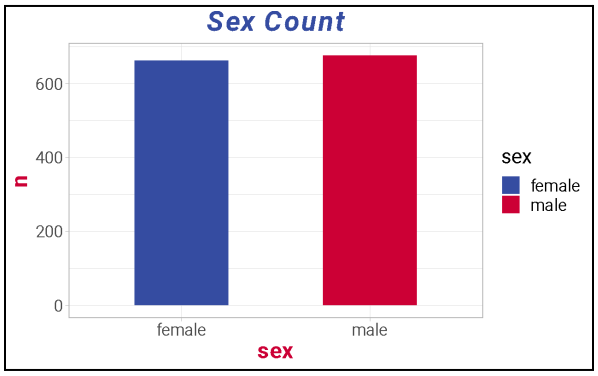
First 6 rows of the data:

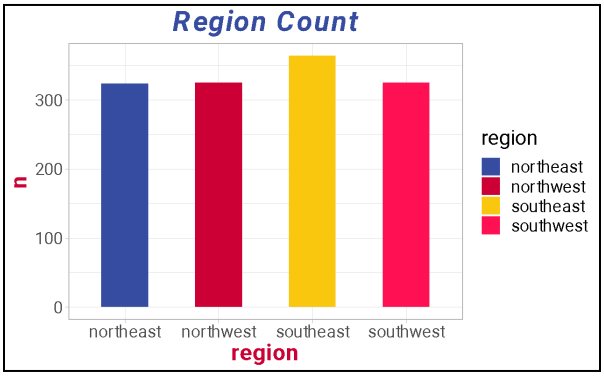
age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622

Exploratory Data Analysis

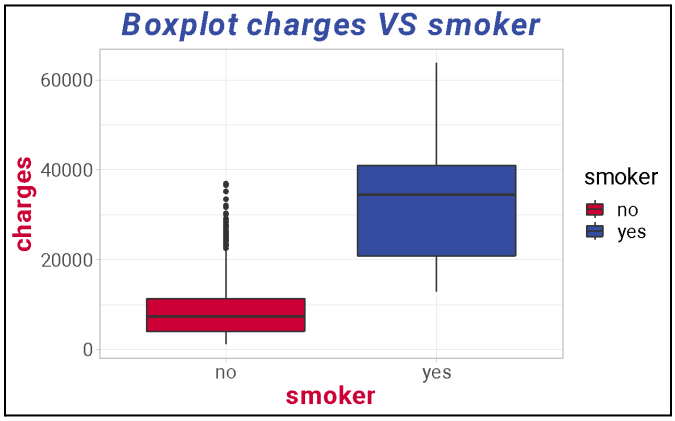
We will initially assess the distributions of the explanatory variables. We find that counts for sex and region are roughly balanced between their respective classes. We also find that the distribution of bmi observations is roughly normal, centered at approximately 30. For age, however, we find that 18- and 19-year-olds are substantially over-represented in the data (by roughly a factor of 2), the distribution of children has a heavy right skew, and non-smokers outnumber smokers by roughly 4 to 1.

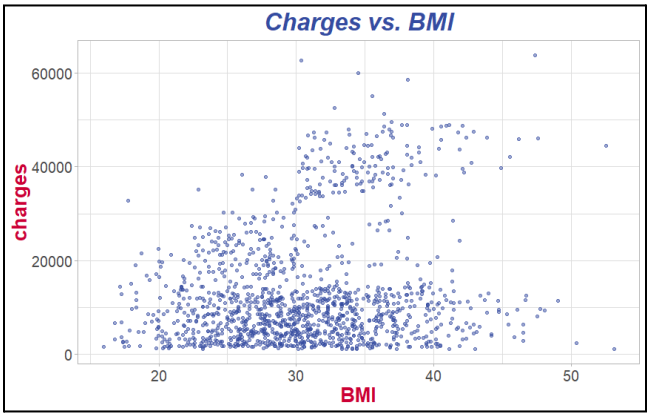




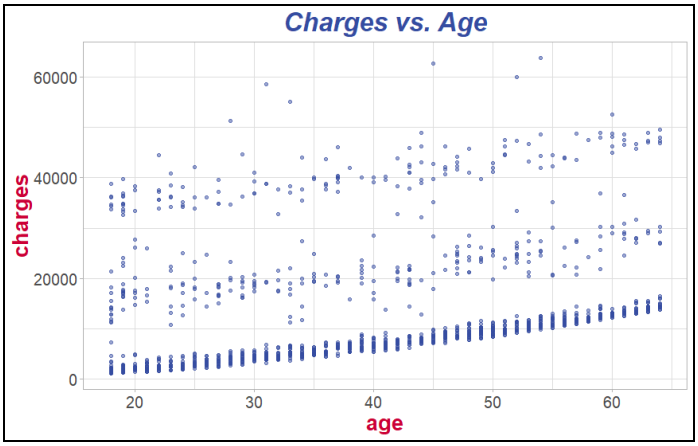


Before we make any changes to the data or begin any modeling, we will first split the data into a training set (80% of the data), a test set (10% of the data), and a validation set (10% of the data). We will now examine the training data for covariance among the variables. We observe that the distribution of charges for smokers is substantially higher than that of non-smokers. We also find that the highest charges are mostly reserved to observations with higher bmi values.



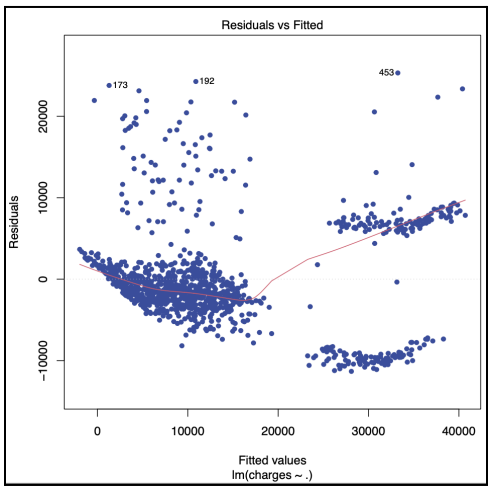


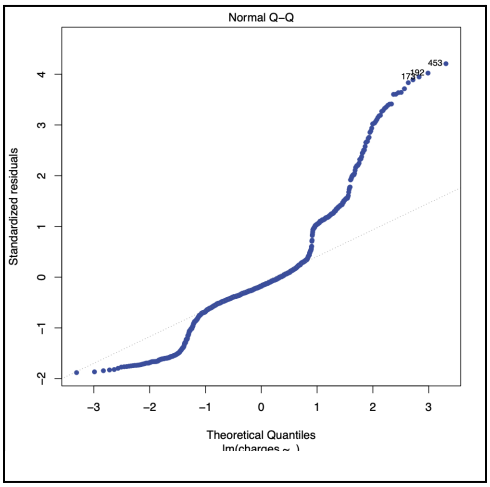
These variables will play a significant role in our analysis. One other interesting pattern of covariance is that of age and charges. There is a plainly visible “banding” or “echo” pattern among the charges. That is, they appear to be grouped into three distinct tiers. Each of these increases with age; the first (lowest) tier does so quite strictly. Somewhat suspiciously, the charges reported in the data often have 3 decimals of precision (atypical for US currency). The unusual charges, as well as this non-random pattern in the data warrants further investigation. With zero documentation accompanying the data, we can only speculate about what underlying processes may explain this phenomenon. We will incidentally revisit this when we try to capture effects like these when looking at interaction effects later.



Simple Linear Regression Model

We will first attempt a simple linear regression model on the training data as-is. This is a relatively straightforward process. We will use the results as a benchmark that we refer back to later after some modifications to the model. When we fit a linear model to the training data, we find that the model's performance leaves a lot to be desired. We calculate an adjusted R^2 value of 0.74 and a root mean

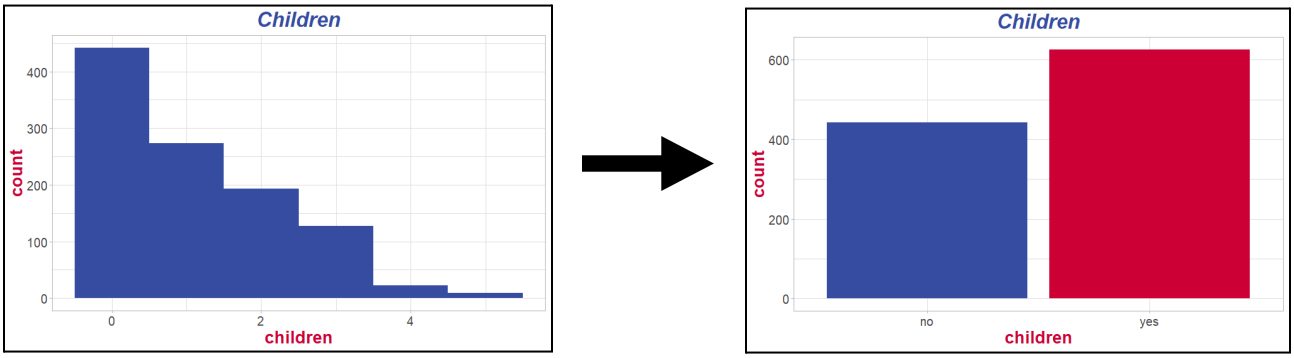




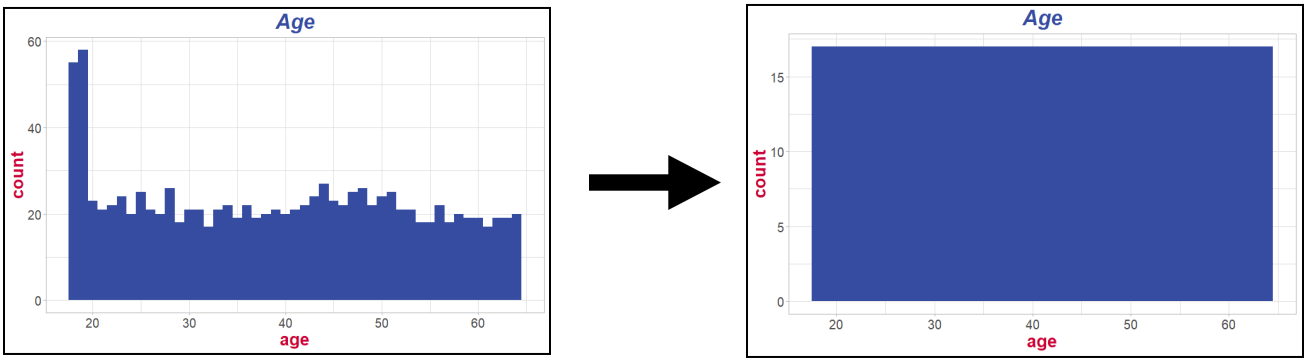
squared error (RMSE) of 6029. We can also observe that our basic model substantially violates our assumptions of and normality assumptions.

Addressing Class Imbalances

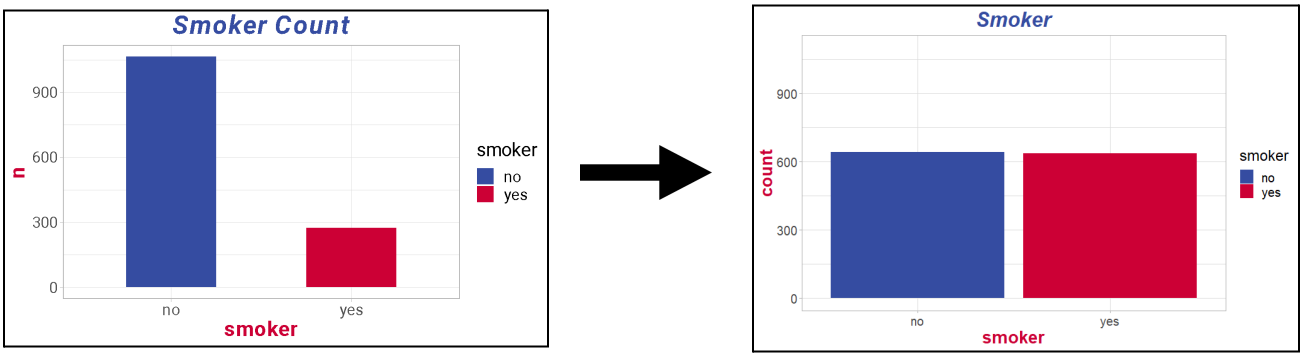
We will attempt to improve this model by addressing class imbalances and considering interaction effects. First, we will collapse the children variable into a binary indicator. That is, if the number of children is greater than 0, we will change the observation to "yes," and if the number of children is equal to 0, we will change the observation to "no." This is done in large part due to the severity of the class imbalance. There are a mere 9 observations with 5 children. Attempting to resample such a severe class imbalance (<0.1%) would do more harm than good. This decision was also made due to the relatively low importance of the children variable in the basic linear model. Because we want to discard this additional information (the precise number of children) for all of the data, we repeat this process on the test set and validation set.



Next, we will correct the class imbalance in the age class. Because we have only two (out of 47) age groups that are overrepresented, we will balance this class using downsampling: observations are removed at random from over-represented ages until the number of observations matches the number of observations for the age with the fewest observations. The result is that many observations are removed from the 18 and 19 ages, while only a few observations are removed from the other ages.

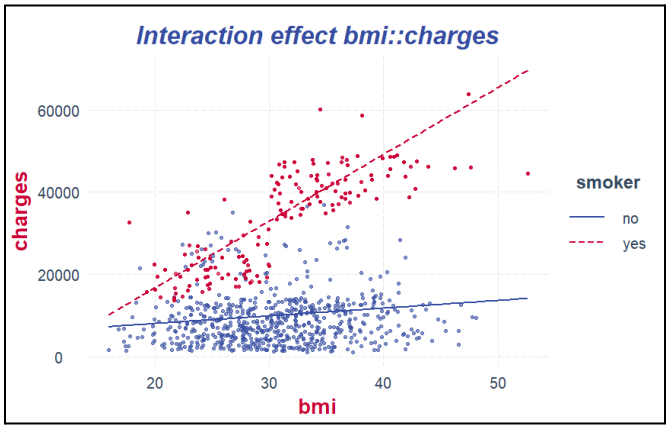


Lastly, we will address the imbalanced ratio of smokers and non-smokers. Because the class imbalance is less severe for this feature, we will use resampling: observations of smokers are resampled at random and added back into the training set as 'new' observations until the number of smokers equals the number of non-smokers.



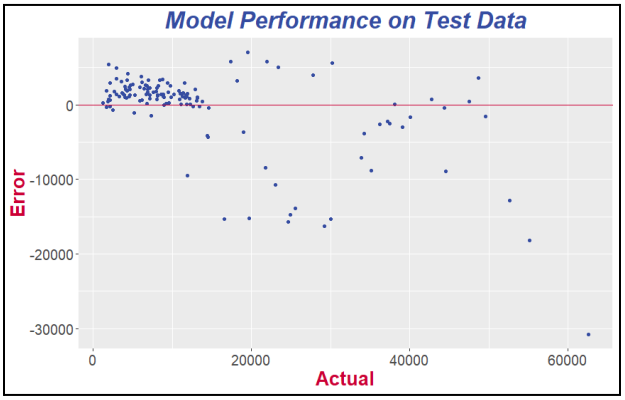
Interaction Effects

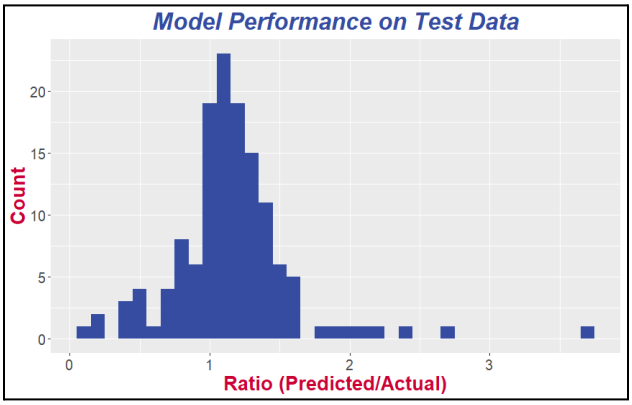
To further improve the linear model, we will now consider interaction effects. We calculate an intermediate model with all combinations of second order interactions, then reduce the model by removing the interaction variables which are insignificant. We determine there are nine significant interaction effects: age::sex, age::children, age::region, sex::smoker, sex::region, bmi::children, bmi::smoker, bmi::region, and children::region. We now create a new model, including these effects, and fit it to the adjusted training data.



Reduced Interaction Model

We fit the reduced model to the training data and calculate an adjusted R^2 value of 0.89 and an RMSE of 5100 - both metrics showing an improvement over the basic linear model's performance. As a cursory check for overfitting, we also calculate these values for the basic model and reduced interaction model where we again observe an improvement in both metrics.





That's not to say that the model is perfect, however. The model's performance tends to underestimate charges in the low range, while performance degrades (and variability increases) substantially for observations as charges increase. These opposing faults somewhat offset each other as we observe a predicted-to-actual plot of the model's performance is roughly normal and centered *near* 1. There are perhaps a few more ways in which this model could be improved, but the gains in performance would likely be marginal. Attempts to further improve a linear model's accuracy may be stymied by limitations of the data itself. More importantly though, a linear model may not be the most appropriate method for approaching this problem (as somewhat evidenced by the aggressive violations of the model assumptions). We will now consider other model types to predict charges from the given data.

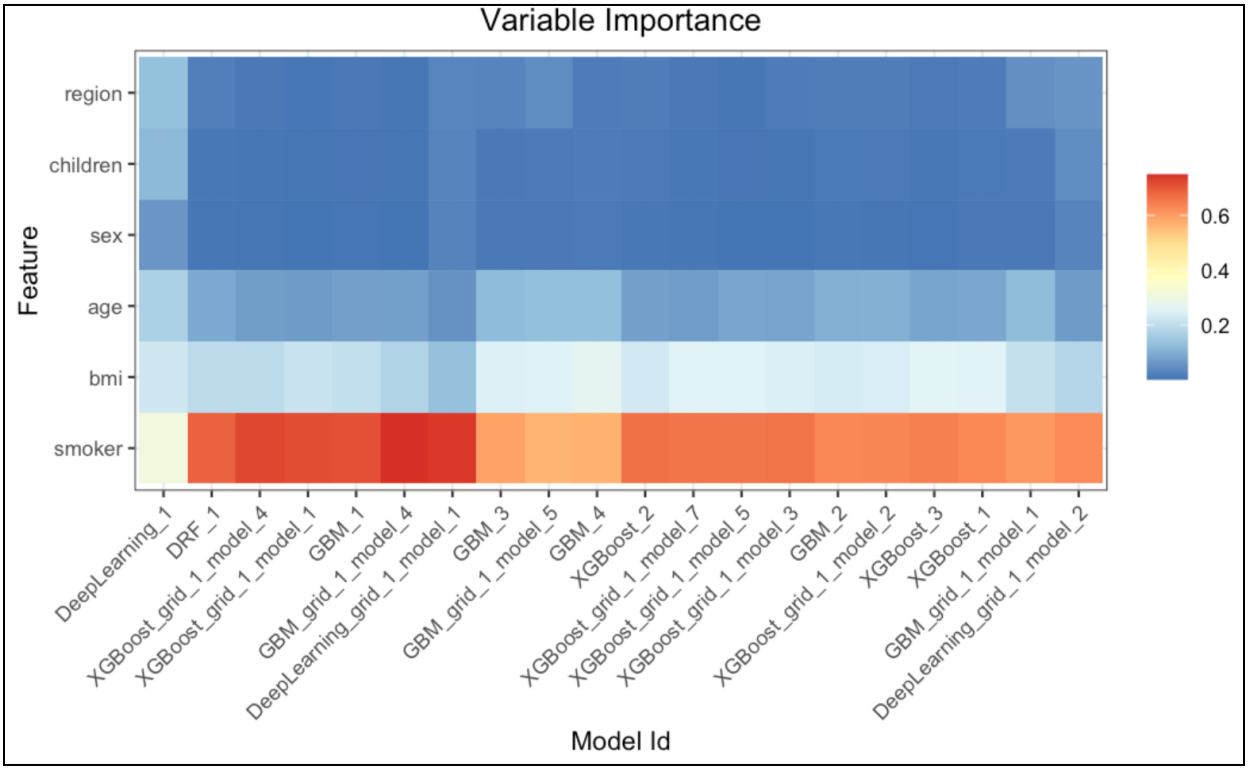
h2o Auto-Modeling

After the basic linear models, other advanced models are searched to make better predictions on the test set. h2o autoML is a user-friendly wrapper function to train and tune a lot of cutting-edge machine learning algorithms and deep learning models. By using the same training, validation, and test set, and running all possible machine learning model, the results are as following:

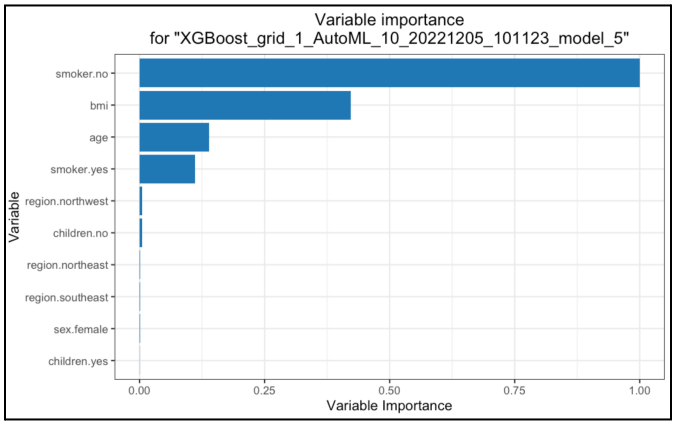
	model_id	rmse	mse	mae	rmsle	mean_residual_deviance
1	XGBoost_grid_1_AutoML_10_20221205_101123_mod...	5116.826	26181910	2766.775	0.4450060	26181910
2	XGBoost_grid_1_AutoML_10_20221205_101123_mod...	5170.155	26730504	2835.416	0.4624419	26730504
3	GBM_grid_1_AutoML_10_20221205_101123_model_4	5215.750	27204048	3087.991	0.4406351	27204048
4	GBM_1_AutoML_10_20221205_101123	5291.722	28002322	2926.122	0.4468681	28002322
5	XGBoost_grid_1_AutoML_10_20221205_101123_mod...	5432.403	29511004	3211.885	0.4799171	29511004
6	XGBoost_grid_1_AutoML_10_20221205_101123_mod...	5460.100	29812690	3125.481	0.6697360	29812690
7	StackedEnsemble_AllModels_1_AutoML_10_20221205...	5485.736	30093301	3197.888	0.4786006	30093301
8	XGBoost_3_AutoML_10_20221205_101123	5511.491	30376538	3254.756	0.5296563	30376538
9	StackedEnsemble_BestOfFamily_1_AutoML_10_20221...	5554.349	30850797	3017.943	0.4682170	30850797
10	XGBoost_grid_1_AutoML_10_20221205_101123_mod...	5558.332	30895050	3255.699	*NA*	30895050
11	DeepLearning_1_AutoML_10_20221205_101123	5576.298	31095104	4005.506	0.5182924	31095104
12	GBM_2_AutoML_10_20221205_101123	5582.627	31165721	3528.015	0.5980802	31165721
13	DeepLearning_grid_1_AutoML_10_20221205_101123...	5614.633	31524100	3270.893	0.4273912	31524100
14	XGBoost_1_AutoML_10_20221205_101123	5640.868	31819393	3554.538	*NA*	31819393
15	DRF_1_AutoML_10_20221205_101123	5687.767	32350689	3323.658	0.4848055	32350689
16	GBM_grid_1_AutoML_10_20221205_101123_model_5	5719.135	32708506	3481.174	0.4848750	32708506
17	GBM_grid_1_AutoML_10_20221205_101123_model_1	5756.339	33135438	3474.058	0.4910644	33135438
18	XGBoost_2_AutoML_10_20221205_101123	5772.284	33319265	3713.435	0.6284526	33319265
19	GBM_4_AutoML_10_20221205_101123	5772.707	33324146	3752.568	0.5485455	33324146
20	GBM_3_AutoML_10_20221205_101123	5848.088	34200138	3758.589	*NA*	34200138
21	DeepLearning_grid_1_AutoML_10_20221205_101123...	5866.296	34413428	3128.297	0.4090168	34413428
22	XGBoost_grid_1_AutoML_10_20221205_101123_mod...	5904.629	34864645	3652.983	*NA*	34864645
23	GBM_grid_1_AutoML_10_20221205_101123_model_3	5934.411	35217233	3700.904	0.5208118	35217233
24	GBM_5_AutoML_10_20221205_101123	6012.663	36152115	3846.900	0.5190894	36152115
25	XGBoost_grid_1_AutoML_10_20221205_101123_mod...	6141.902	37722963	4111.783	0.6144804	37722963
26	GBM_grid_1_AutoML_10_20221205_101123_model_2	6424.870	41278950	4473.232	0.5633821	41278950
27	XRT_1_AutoML_10_20221205_101123	8216.696	67514085	6870.960	0.7553406	67514085
28	DeepLearning_grid_3_AutoML_10_20221205_101123...	9552.470	91249686	8604.414	0.9199718	91249686
29	DeepLearning_grid_2_AutoML_10_20221205_101123...	9841.567	96856444	9037.470	0.9226905	96856444
30	DeepLearning_grid_2_AutoML_10_20221205_101123...	9855.466	97130219	9028.305	0.9699905	97130219
31	DeepLearning_grid_3_AutoML_10_20221205_101123...	9938.857	98780876	9146.109	0.9351571	98780876
32	GLM_1_AutoML_10_20221205_101123	14648.506	214578721	13127.164	1.1573980	214578721

Thirty-two models are run and h2o autoML is running grid search and hyperparameter tuning automatically. This list is ranked by the lowest RMSE on the test set. It is obvious to find that XGBoost_grid1_AutoML is the best one with 5116 RMSE on the test set. From this list, the top 10 models are all tree-based models. The dataset contains the quantitative and qualitative variables. In this case, there is no surprise that tree-based models perform the best among all machine learning models, which are also better than the deep learning model.

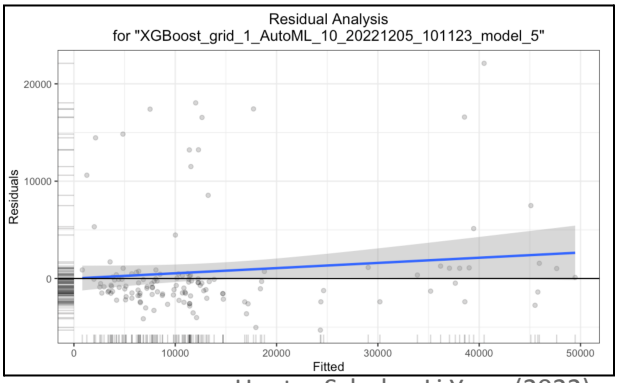
The following plots shows variable importance for some of models. It is obvious to see that smoker is the most significant variable impacting insurance charges. The bmi is the second important variable to impact charges regardless of model types. However, the deep learning model doesn't give the smoker variable too much importance, an assessment that is quite different from the other models and EDA results. This means deep learning models are very strong in image analysis but less effective at predicting charges like this.



This is variable importance for the best model. As we can see, the smoker is the most significant variable; it is twice as significant as bmi, the second most significant variable. Region, sex and children don't impact the charges by much, which is consistent with our EDA analysis.



This figure shows the residual analysis of the best model, XGboost model. As long as there is no obvious pattern in this graph, the residuals are acceptable. There is a fitted blue line and confidence interval around it which shows that there is a slightly increased pattern in this residual plot; however, the slope is not steep. Instead, the slope is flat and close to the horizontal line. This indicates that the residuals are acceptable.



Conclusions

This is a small dataset. More accurate data would significantly increase the model performance on the test set. The number of non-smokers is much more than that of smokers, which may be impacting the model's effectiveness. Smoking will significantly increase insurance charges. Sex, children and region have almost no effects on charges. Additional

variables such as alcohol consumption could increase model accuracy. Basic linear models substantially violate the model assumptions, and may not be an appropriate modeling technique for this data set. Similarly, the Deep Learning model is not very suitable in this case. The tree-based (47 trees) XGBoost model has the lowest RMSE on the test set by auto tuning its parameters.

Actionable Insights

End-users could utilize this model in several ways. Insurance companies (or perhaps employers) could make a mobile app or website for customers to input their bmi, whether or not they smoke, and other key factors to dynamically predict their insurance charges next year with this best XGboost model. Additionally, insurance companies could use such a model to evaluate new customers. When a new customer considers the company's insurance plan, the company could use the model to give interval prediction of charges for the customer. Companies could internally make predictions to decrease the costs passed to the customer in order to compete with other health insurance companies.