Machine Learning and Predictive Analytics in the Context of Pricing Insurance Policies

DBA3803 Predictive Analytics in Business

November 26th 2023

Group 19

Michail Eric Marechal (e1120253), Zofia Maria Sosnowska (e1120255), Sofia Roche Vidaurre (e1120271)

Intro:

In the realm of insurance analysis, the importance of leveraging comprehensive datasets and employing robust models cannot be overstated. This report delves into the intricate process of constructing a predictive model for insurance charges, exploring the nuances of the dataset, the methodology employed, and the performance evaluation of various models. Throughout this process our main focus will be limiting underestimation. For insurance companies it is imperative to not set insurance rates too low as it can lead to huge losses caused by unexpected exposure to risky individuals. However, due to the intense competition and the lack of possibilities to differentiate insurance companies cannot significantly increase their policies' prices otherwise their clients will switch to another contractor. For this reason, we will be limiting underestimation only to a certain extent, to avoid constant excessive overestimation. In the conclusion we will suggest which predictive model is best suited for this dataset and for predicting appropriate insurance charges while ensuring interpretability.

Data:

The data set to be used throughout this paper and therefore in the constructed model is based on seven different covariates: age, sex, BMI, (number of) children, smoking habits, region, and finally the insurance charges. Within that said there was a need for encoding the different nonnumerical values, therefore variables sex and smoker have been denoted in a binomial form, with 1 meaning male and 0 a female for the former of the two variables, and 1 indicating a smoker and 0 a non-smoker for the latter. Another categorical variable within the dataset, namely region, has been transformed into a dummy variable to ensure better usability of the data within the machine learning models.

The dataset comprises 1338 observations, gathered among clients of the health insurance industry. Age wise the highest percentage of the sample was made up for people at the age of 18 and 19, while the differences between other people within the dataset were mostly uniform and stable across all age groups. In terms of the sex, the data was split almost evenly between men and women. The number of children referring to the number of dependents covered by the insurance seemed to be equal to zero in the majority of the observations, with progressively fewer cases with the increasing number of dependents.

The most significant difficulty connected with using big data sets in one's model is the risk of overfitting. Overfitting occurs when a large number of data points fits the training test too closely and therefore more often than not leads to a failure in generalising to the test set. The simplest way to measure whether the issue of overfitting is one to be seriously worried about is by checking the rule of thumb. In this particular case multiplying the number of covariates by a ten yields 700 which is lower than the number

of observations. This means that one should not be too concerned with the issue of overfitting for regression models (Lasso, Ridge and Quantile Regression).

Procedure & Loss:

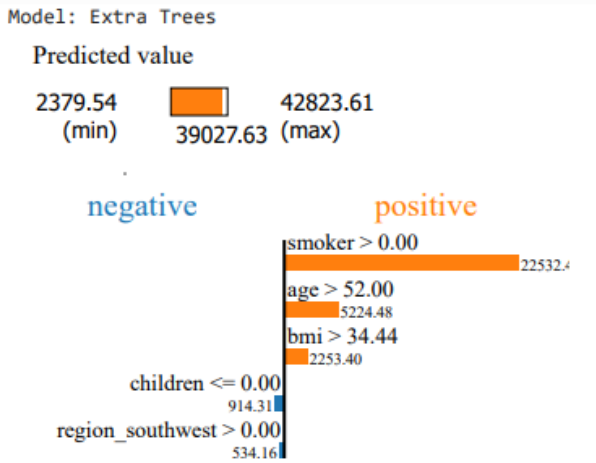
The journey into model construction involves the utilisation of various techniques, including LASSO, Ridge, Random Trees, Quantile Regression and Gradient Boosting. The first three models were trained using L2 as their loss function, whereas the two last ones used a Pinball loss of 0,75. The models using L2 as their loss function were then adjusted by 102% to reflect our willingness to mitigate the risk of underestimating. Then, based on our research, we will employ LIME, a tool adept at dissecting model predictions at an individual level. This in-depth analysis, focused on specific observations, should unveil critical insights into the intricate relationship of the variables influencing insurance charges. Barriers and divisions will be meticulously established for variables such as smoking, age, number of children, BMI, region, and sex. Finally, all models will be compared and the best model will be chosen as a recommendation for predicting future insurance charges.

Interpretability:

In order to look into the interpretability of the chosen model LIME is one of the tools that can potentially be of use. This particular instrument focuses on evaluating and explaining the predictions of the model based on individual observations rather than the entirety of the dataset. The so-called explainable model is predominantly used in order to gain a deeper understanding of the importance of variables for the individual prediction of interest.

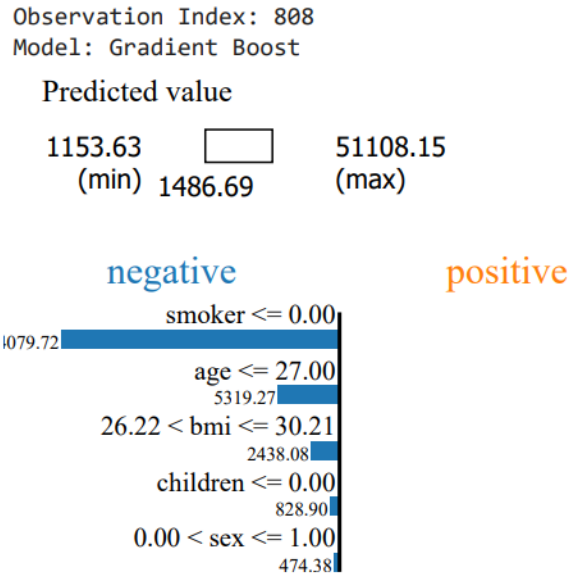
The main advantage of using this method is that it is suitable for any kind of Machine Learning models, all models explored throughout this paper included. As far as the relevance of this practice goes, in the business world interpretability of the model's output is crucial for making informed data-driven decisions. This is due to the fact that more often than not the decision-makers come from a non-technical background, hence are in need of an easier to understand output.

In the analysis observations with the following indexes were used: 808, 175, 286. For each of the observations LIME was performed on three structures namely Gradient Boot, Extra Trees and Random Forest. The output of the analysis has presented us with a total of 9 graphs showcasing detailed information about the influence of particular variables on the price/premium prediction of the models with respect to particular observations and the structure used.



The following barriers were imposed on the variables. Firstly, the value of 0 of the smoker variable will have a negative impact on the predicted value, while the failure of 1 will increase the prediction. Next there's the age variable that has been split at the age of 52, where each value below the threshold influences the prediction in a negative way. The division in the variable denoting the number of children covered in the insurance package is set at 0 given the frequency of that particular observation within the dataset. Similar division points were also established for the variables BMI, region and sex. [1]

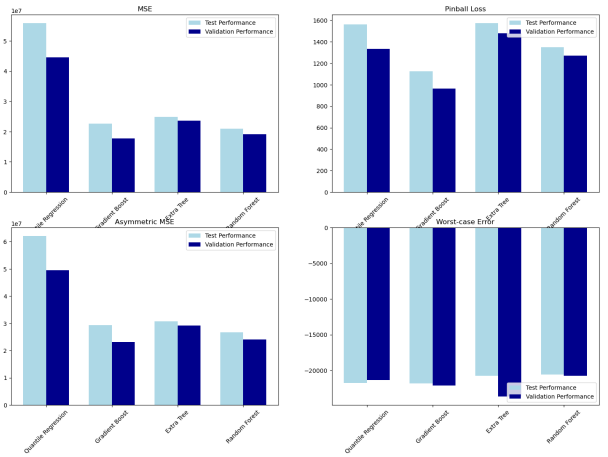
From the analysis we can infer that the main price/premium determining factor is whether or not a person purchasing the insurance is a smoker. Therefore the highest weight will be assigned to this particular variable within the dataset. Another quite important factor seems to be the age variable, as it is the one being considered in all kinds of models and for all observations of interest. Similarly, the variables BMI and (number of) children appear in all considerations. However, there's one main distinction between the Gradient Boost, Extra Trees and Random Forest. In the former structure of the model, the fifth variable that has significant impact on the prediction is the variable sex, while in the latter two, the last significant variable is the region.[2]



[1] The LIME output for the observation 175 for the Extra Tree model.
[2] The LIME output for the observation 808 for the Gradient Boost model.

Performance comparison:

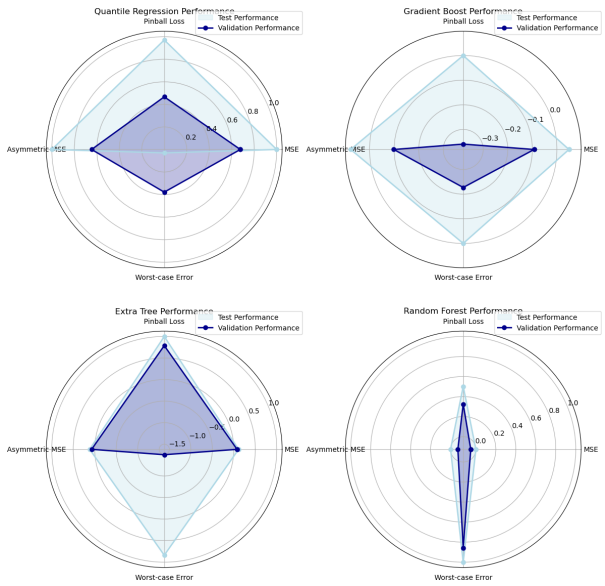
Now we move on to the final performance of our trained models and finally reveal our test dataset, which has been left purposefully unused. After training our various models Quantile regression and Gradient boost which both used a 0,75 pinball loss as their loss function, the latter being tuned by using a random search grid technique, with 20 iterations, to select hyperparameters for: number of estimators, learning rate, maximum depth and minimum number of samples per leaf. Those parameters were chosen to further refine the model in order to get better results, but at the same time, the objective was to avoid overfitting at the expense of validation performance. The two last models were Extra Tree and Random Forest. Both models were trained using an L2 loss and as required by the business context we wanted to limit the quantity of underestimation as those can result in unexpected additional risk carried by the insurance company. Therefore, as L2 does not discriminate between under or overestimation, we manually adjusted by 102% to get the desired outcome. Additionally, those two models were tuned for the following hyperparameters using a random search grid with 20 iterations: number of estimators, maximum depth, minimum samples per leaf and the maximum number of features considered at each split. In accordance with what was said previously, the aim was to improve the model, but make sure at the same time to mitigate overfitting as much as possible.



Based on the following results we can see that for the first three performance metrics (Mean Square Errors, 0,75 pinball loss, Asymmetric Mean Square Errors which penalises underestimation and Worst-case error) our models perform worse on the test set compared to the validation set. This is logical as our model was trained using the validation set, therefore, the model was trained to accommodate the underlying data. However, we implemented numerous techniques to mitigate the risk of overfitting, as we didn't want to have models unable to generalise to other datasets. Although our performance for the test set was worse, as explained above, we can still observe not a very large difference between the test and validation set performance suggesting that we were able to mitigate successfully the risk of overfitting.

Subsequently, what is worth observing is that all the models, except quantile regression, performed better for the test set compared to the validation set on the performance metric of Worse-case analysis. This is a good result as those extreme individuals who are largely underestimated by linear models can be the

source of excessive risk taking without any financial reward, therefore, the fact that our models performed well on that metric is an excellent result. Finally, we were able to observe that linear models, in this selection quantile regression, almost always consistently perform the worst for the different measured metrics, this suggests and confirms the non-linearity in our data and thus the reason why they are unable to accommodate, conversely to tree-based models.



Lastly, if we look at the radar charts, we can see that most validation performances are surrounded by test performance suggesting a better performance on the validation set. For all the metrics, except worse-case analysis, as we move further from the origin the worse is the performance, whereas for worse-case analysis it is the opposite, as a higher score suggests less underestimation error. It offers a better visual way to analyse how proportionally the models perform on each performance metric. Please be careful that the scales are not the same for all the radar charts.

Conclusion:

In conclusion, we can see that overall Gradient Boost and Random Forest are the two models that standout. Both models performed best on all the metrics and notably Random Forest on worst-case analysis, where it was able to accurately handle extreme underestimation predictions. A reason for both models' good performance was also their ability to accommodate the non-linearity in the data and their predictions. Gradient Boost performs significantly better on Pinball loss, as it was expected, as it was trained for it and Random Forest performs better on extreme cases, as expected as L2 loss penalises very divergent predictions. Based on these results we advise insurance companies to make use of the Gradient Boost in usual times, as it is efficient at obtaining low pinball loss and thus aggregate underestimation, however, in turbulent times insurance companies might want to switch to our Random Forest model, as in those situations unexpected risk exposure from largely underestimated individuals can cause the company a lot of troubles.