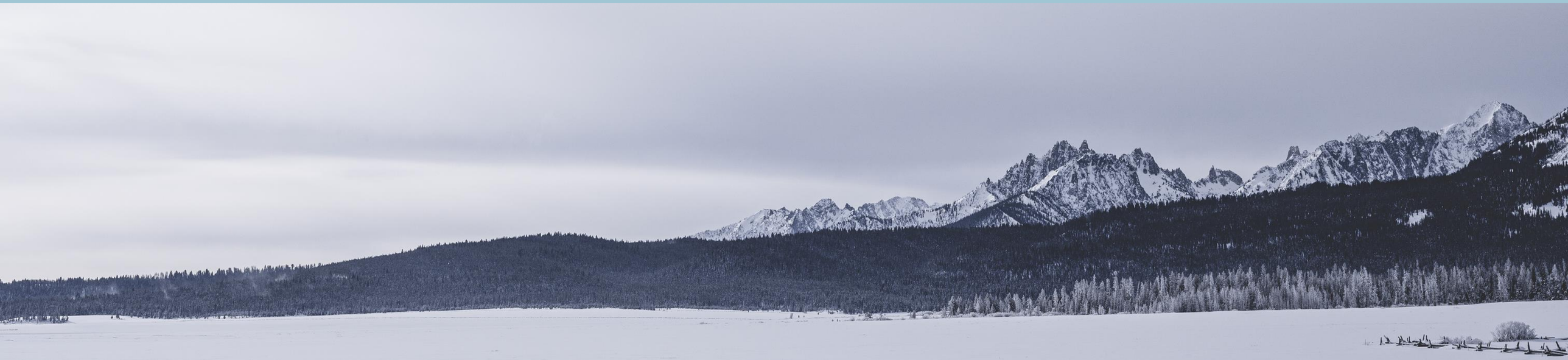
Interpretability: Peeking inside the Black Box using ICE and PD plots



Jazel Jayme, MSDS 2022

Learning Goals

- To have a high-level idea of Partial Dependence (PD) and Individual Conditional Expectation (ICE) plots

- To use PD and ICE plots in interpreting the results from trained black box model

Interpretability vs Accuracy

Classical Models

- Can be interpreted.
- Give information on how X and Y are related

Linear Regression
Logistic regressions

Black box Models

- Cannot be interpreted.
- Can model complex data

Random Forest
XGBoost
Neural Networks

Partial Dependence Plot

- Answers the question: Given a feature *S*, what is its **average marginal effect** on the predicted outcome?
- Can capture more complex patterns from your data, and they can be used with any model.

Individual Conditional Expectations

- ICE plots are the equivalent to PDPs for **individual data instances**.
- A PDP is the average of the lines of an ICE plot.

Methodology

- To train a model that generates the charges more accurately.

Dataset

- 1,338 medical insurance records
- Patient’s info: age, gender, region, BMI, no. of children, smoker flag
- Target: charges

Preprocessing
And EDA

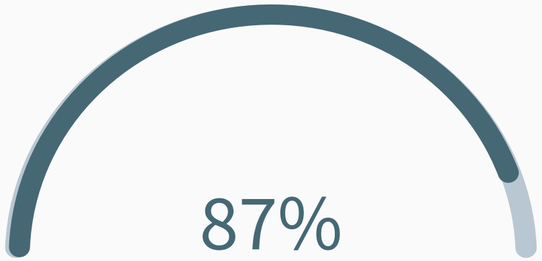
- One-Hot-Encoding for categorical variables
- Univariate and bivariate analyses

ML models

- 7-in-1 Auto ML Regressor
- Compare ML performance

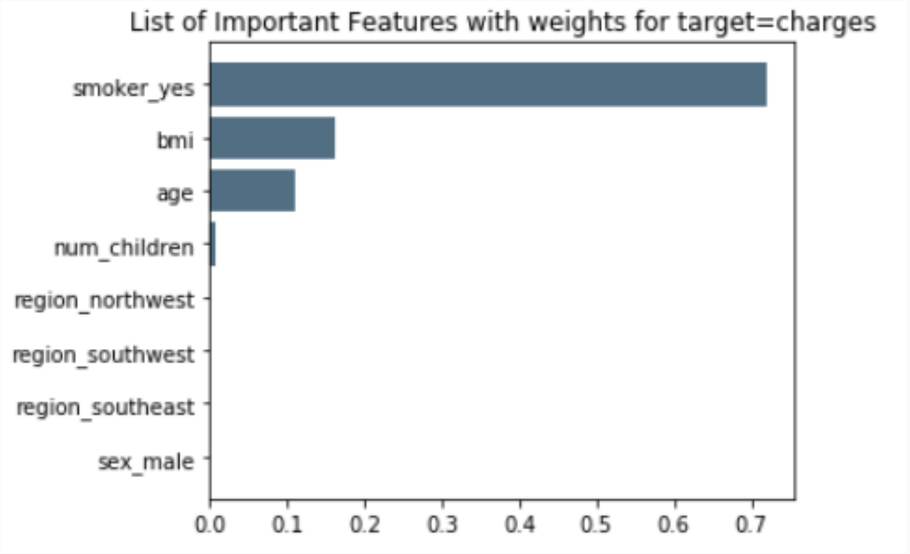
Results and Discussion

Best model is Gradient Boosting!



Accuracy

Top Predictor
Smoker flag



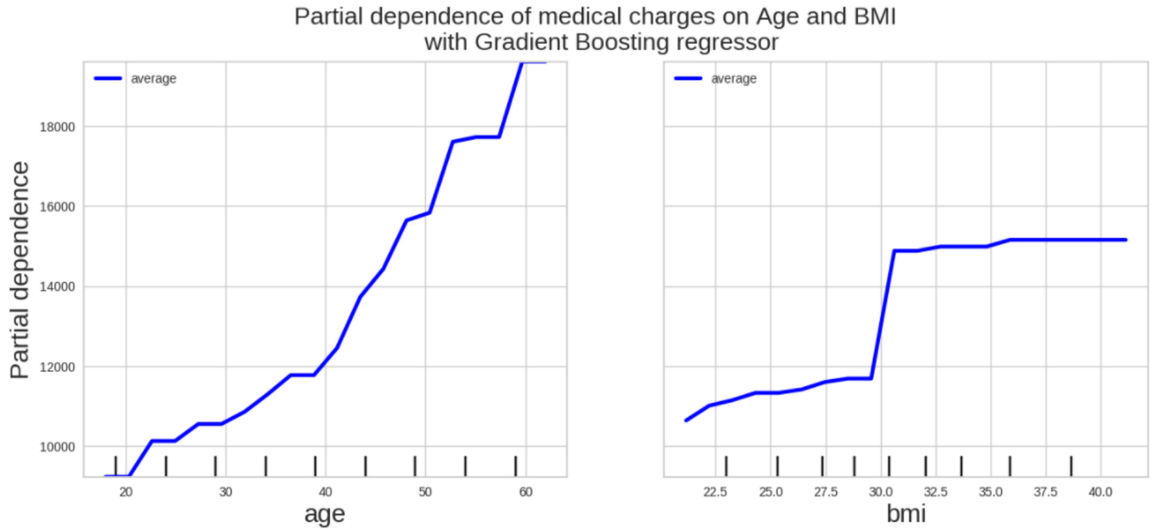
Feature
importances

It does not end there.

How does each feature affects the predicted outcome?

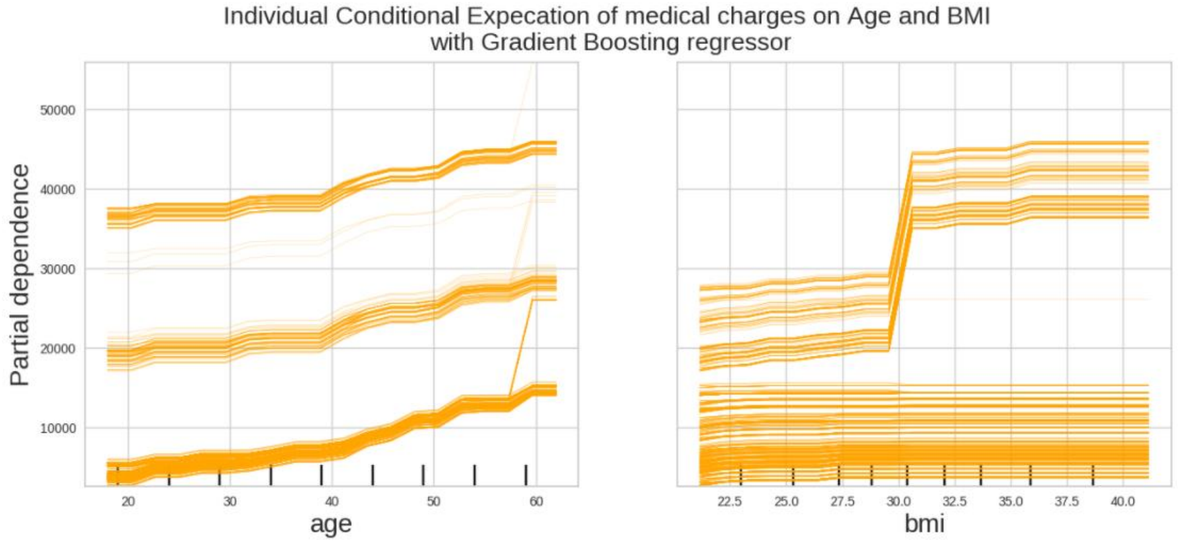
How does BMI affect predicted medical charge?

PD Plot



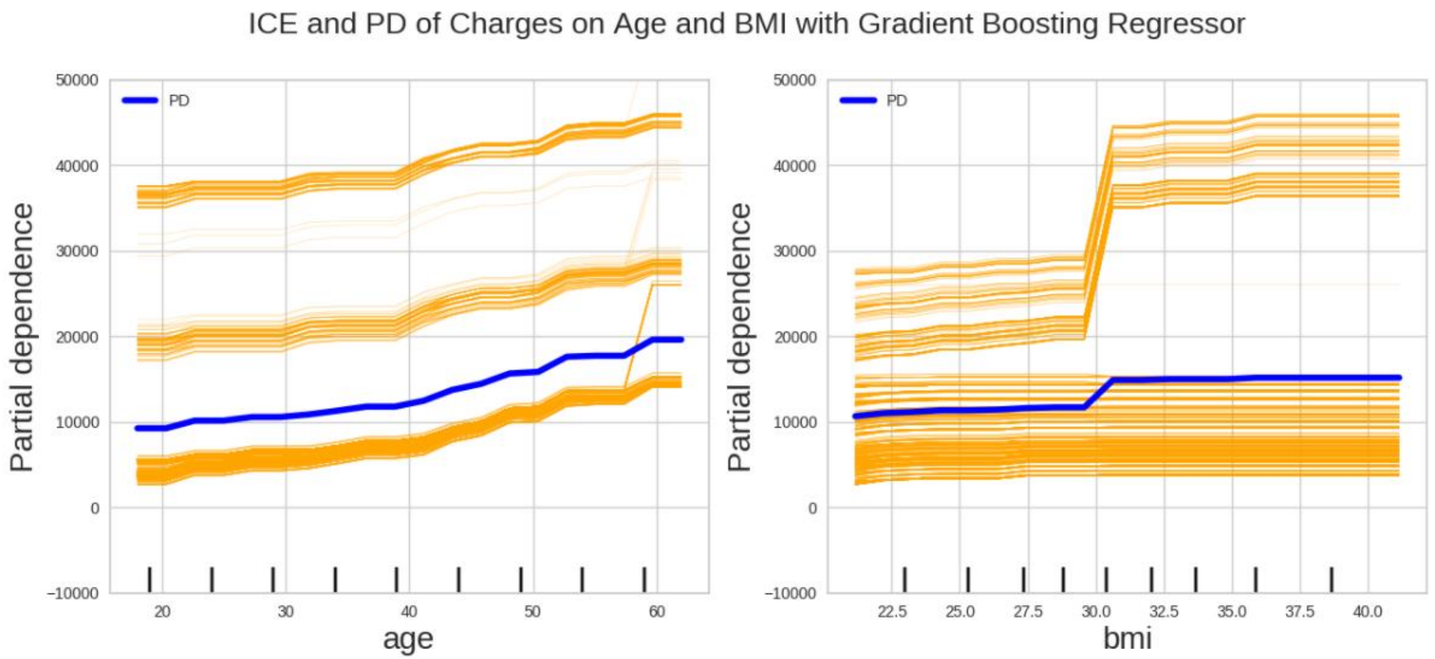
- Older individuals have higher medical charge.
- The charges are low until BMI goes beyond 30.

ICE Plot



- Samples that are clustered represents groups of people with similar behavior or lifestyle (smoker, non-smoker, etc.)

Combined PD and ICE



- PD can cancel positive and negative trends.
- Combining ICE and PD can show the overall behavior of your ML model.

Key Insights

- It is possible to interpret the behavior of complex ML models using PD and ICE plots.
- These tools assist in bridging the gap between you as a data scientist and non-technical folks.


Salamat. ☺

Appendix

	model_name	opt_hyperparameter	val_score	train_score	test_score	run_time	top_predictor	target
0	GradientBoostingRegressor	GradientBoostingRegressor(max_depth=2, n_estim...	0.841	0.8765	0.8721	1.86227	smoker_yes	charges
1	RandomForestRegressor	RandomForestRegressor(max_depth=3, n_estimator...	0.7914	0.8683	0.8686	6.5078	smoker_yes	charges
2	DecisionTreeRegressor	DecisionTreeRegressor(max_depth=2)	0.5802	0.8326	0.8301	0.259257	smoker_yes	charges
3	LinearRegression	LinearRegression(normalize=True)	0.465	0.757	0.7628	0.315259	smoker_yes	charges
4	Ridge	Ridge(alpha=0.001)	0.6615	0.757	0.7628	0.465592	smoker_yes	charges
5	Lasso	Lasso(alpha=100)	0.6631	0.7551	0.7608	0.420692	smoker_yes	charges
6	KNeighborsRegressor	KNeighborsRegressor(n_neighbors=32)	0.1637	0.1587	0.1645	2.65904	None	charges

References

[1] Dassen,T., Hou, N., Kronseder, V. (2020). Introduction to Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE). Retrieved 13 September 2021, from https://compstat-lmu.github.io/iml_methods_limitations/pdp.html
[2] https://i.redd.it/a32nmihyhp061.jpg
[3] https://png.pngitem.com/pimgs/s/614-6141987_cat-reaction-reactioncat-catreaction-meme-confused-tag-a.png
[4] https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQ8fZd-OGbAvWuZyYIB6isMPYOrst9COjq2xg&usqp=CAU