Dental Cost Project

Executive Summary

Not an introduction nor an abstract

Problem (not full statement)

Data source

Model built and feature used

Quality of prediction

Important next steps

\*I have this problem and I find GLM good because …And here is the result.

Problem

Predict dental cost for purpose of setting more accurate premiums. Improve KPI (measure of success)

Something to guide your model choice such as: interpretable by regulators

Missing data, concurrent prediction, overfitting, interaction,

GLM vs. Random Forest

Data

Source – Meps survey

Early obvious adjustments

Summaries and some graphs/tables

Quality Issues: missing occupation

Clean up actions: remove records

Transformation of creating hasIncome

Preliminary investigation of relationships to target

Approach/Method

Partition of data

Why GLM and Random Forest are considered

GLM:

Distribution and link choices

Interaction checked

Drop some variables

Add IsTeenager

Validate via some charts and MSE against training and hold-out

Random Forest

Select mtry via cross validation

Reduce overfitting by changing maxnodes

Check variable importance (may try to remove unimportant variables)

Validate against training and holdout data

Results

Summary of findings

Recommendations:

Interpretability -> GLM

Implement ability -> GLM

Conclusions/Next Step

Bottom line recommendation

Further work

Other interactions

Further RF hyperparameters tuning

Other GLM distribution/links

Other feature selection

Appendices

Data dictionary

Diagnostic graphs/tables that supplemental