Dental Costs Project

Executive Summary

- → Problem (not full statement)
- → Data source
- → Model built and features used
- → Quality of prediction
- → Important next steps

Problem Statement

Predict dental costs for the purpose of setting more accurate premiums. Model should be interpretable, particularly by regulators. Benchmark goal (measure of success)

Data

- → Source—MEPS survey
- → Early, obvious adjustments
- → Summaries and maybe some graphs/tables
- → Quality issues—missing occupation (why?)
- → Cleanup—removed some records with missing values
- → Transformation of creating hasIncome, creation of PCAs
- → Preliminary investigation of relationships to target

Approach/Method

- → Partition of data
- → Why GLM and Random Forests considered
- → GLM
  - → Distribution and link choices
  - → Interaction checked
  - → Drop some variables
  - → Added isTeenager
  - → Validated via some charts and MSE against training and holdout
- → Random Forest
  - → Selected mtry via cross-validation
  - → Reduced overfitting by changing maxnodes
  - → Validated via some charts and MSE against training and holdout
  - → Checked variable importance (might have tried removing the unimportant variables)

Results

- → Summary of findings
- → Recommendation of model to use
  - → Interpretability => GLM
  - → Implementability => GLM
  - → Accuracy => Random Forest (though about the same)

Conclusions/Next steps

- → Bottom line recommendation
- → Further work
  - → Other interactions
  - → Further RF hyperparameter tuning
  - → Other GLM distributions/links
  - → Consider further feature selection

Appendices

- → Data dictionary
- → Diagnostic graphs/tables that are supplemental