Predicting Risk Levels on Life Insurance Policies

By: Richard Mei

Did you know only 40% of U.S. households own individual life insurance? (Data from 2016)

Prudential's Life Assessment Kaggle Challenge

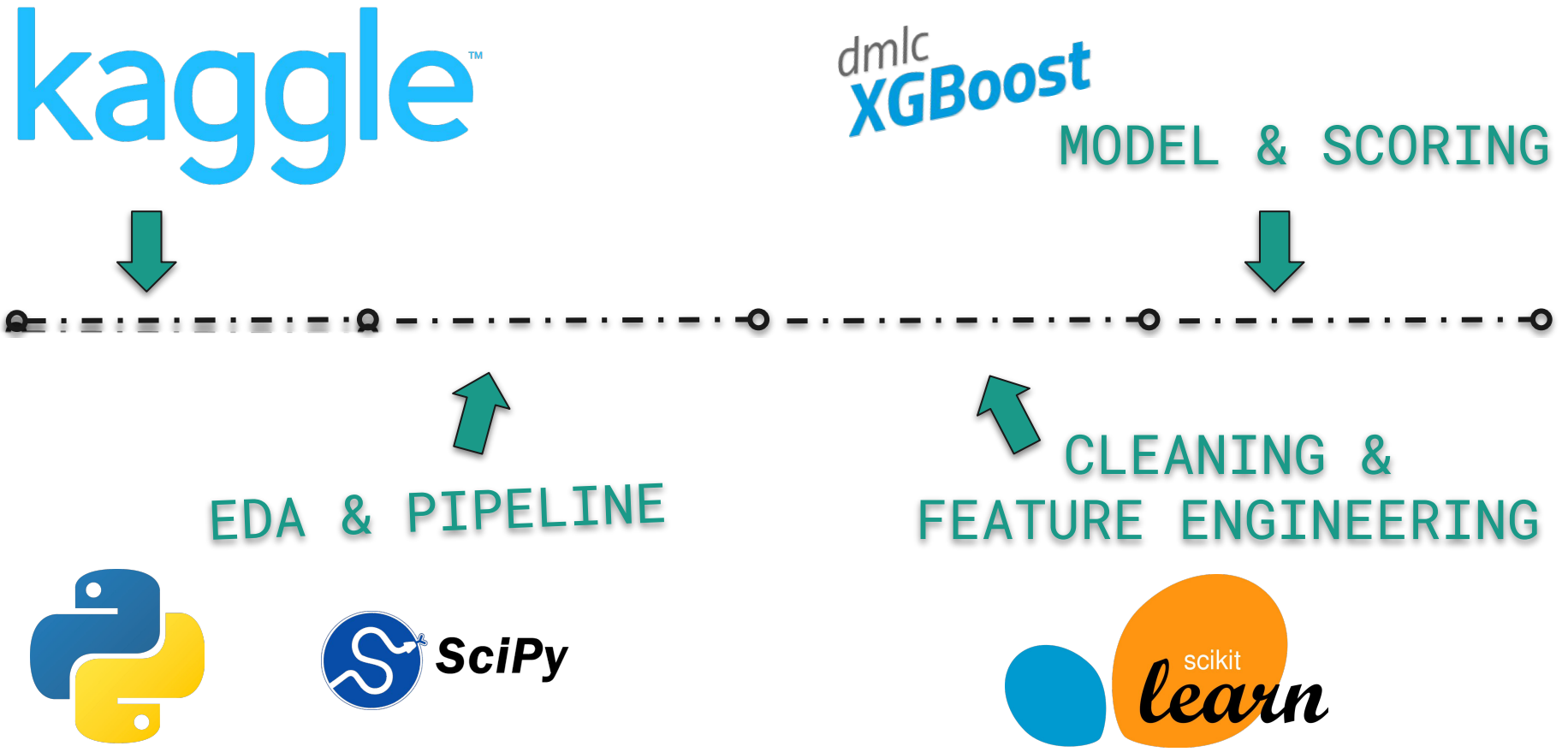
Prudential is a large insurance company in the U.S

Mission: Streamline buying individual life insurance using a predictive model to classify **8 RISK LEVELS**

Challenge: Due to confidentiality, you get no knowledge of what each feature/column mean.

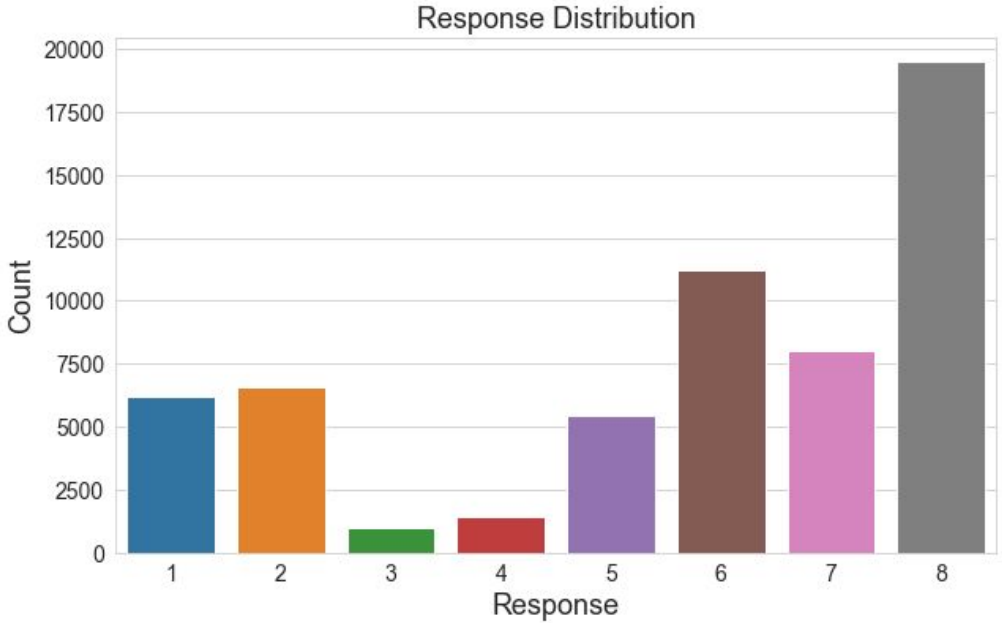
Model's weighted kappa score: 0.56891

(Winning Score: 0.67938)

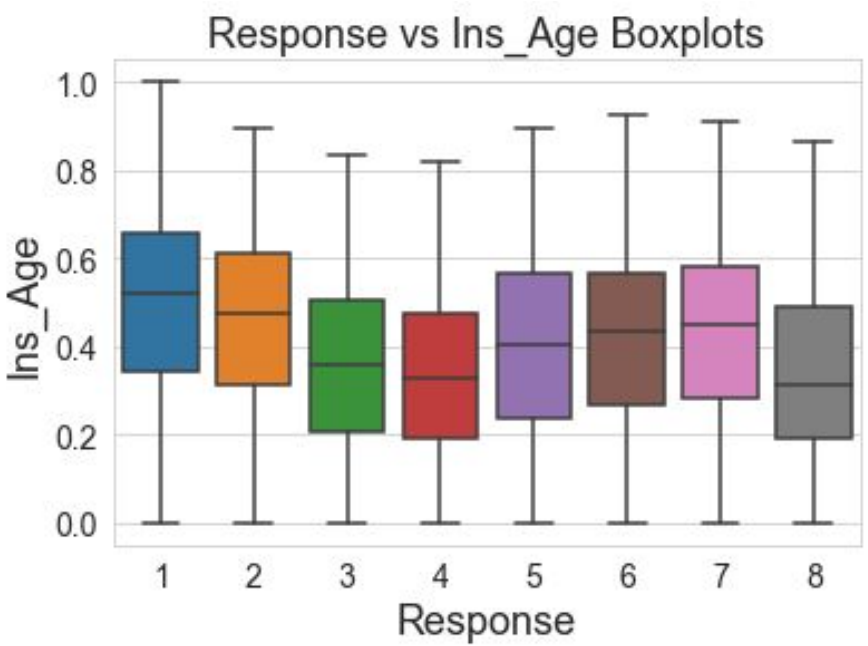

kaggle
dmlc XGBoost
MODEL & SCORING
EDA & PIPELINE
CLEANING & FEATURE ENGINEERING
SciPy
scikit learn

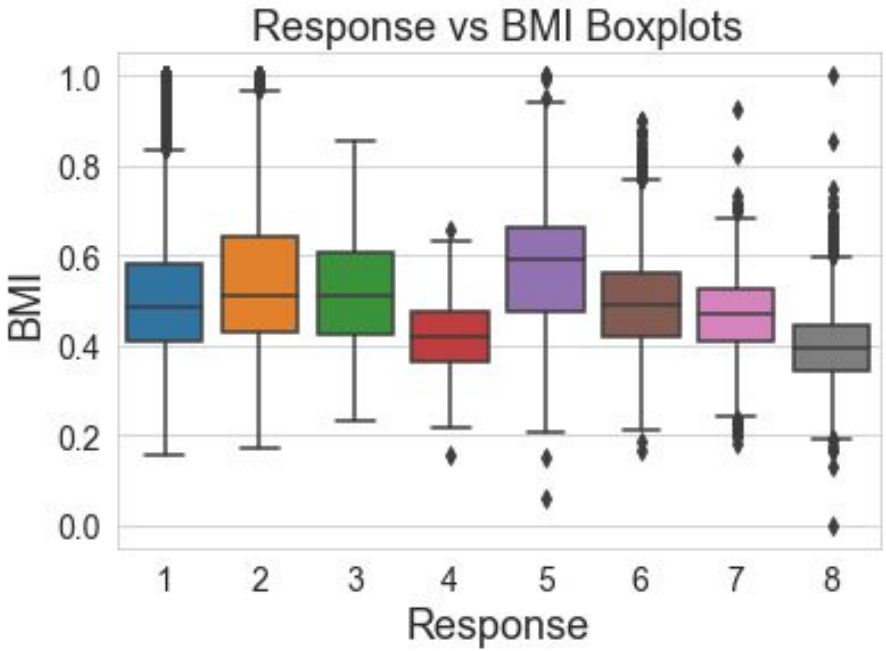
Data

- **Shape:** 59,381 entries by 128 columns
- **Target:** Response (1 to 8)
- **Normalized Features:**
  - Age
  - Height, Weight
  - BMI (cont.*)
  - Employment Info (3 cont., 3 cat.)
  - Insured History (6 cat.)
  - Insurance History (1 con, 8 cat.)
  - Family History (4 cont., 1 cat.)
  - Medical History (5 disc., 36 cat.)
  - Medical Keywords (48 dummy)



*Feature Key:
cont: continuous, cat: categorical, disc: discrete





More EDA: Age and BMI

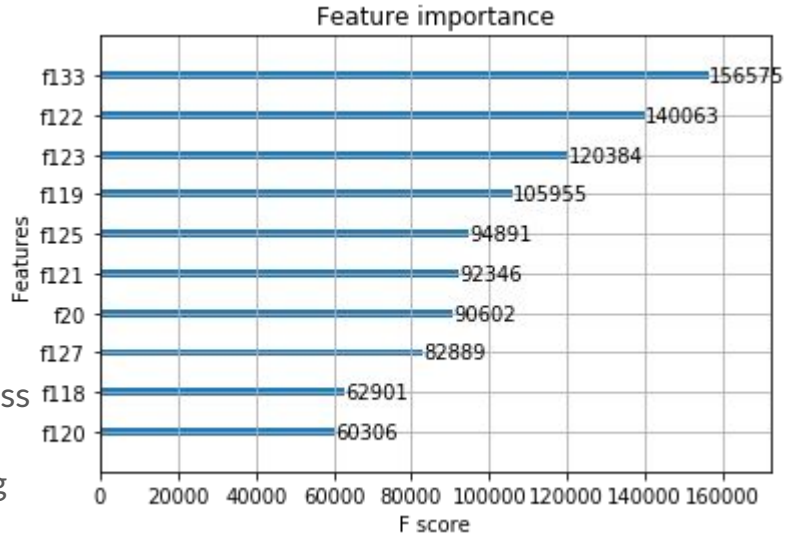
Final Model:

XGBoost Classifier

- N_estimators = 300
- objective = multi:softprob
- eval_metric = ‘merror’

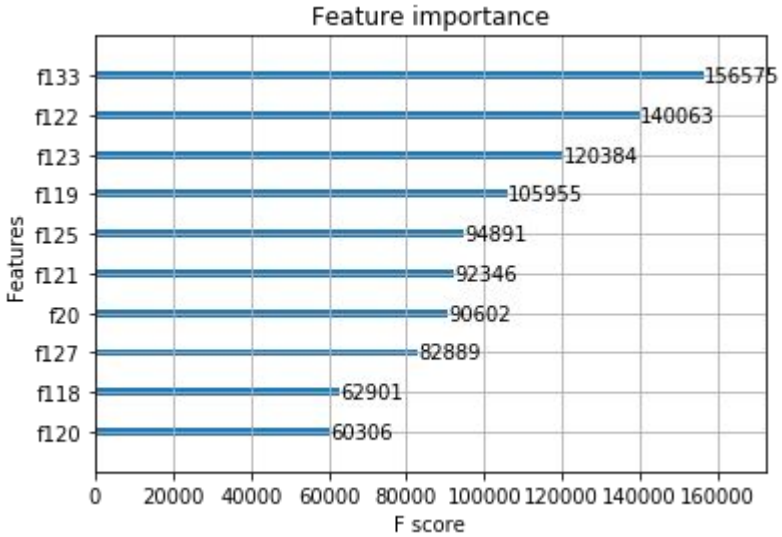
Application of Model

- Train Scores
  - Accuracy: 0.5754
  - F1 weighted: 0.5513
  - Cohen: 0.4571
- Final weight kappa: 0.56891
- Model is more accurate than selecting dominant class
  - Our model can predict
- Figure out which features are important in choosing



What's Next?

- Try using:
  - Downsample (55,000 data points)
  - Linear Regression (Lasso, Ridge)
- Using grid search cross validation in the pipeline to cross validate and also tune for a better model
- Get information on the meaning behind each column for better feature engineering



THE END

Helpful Links

Github: https://github.com/rmei97/mod-3-proj

Kaggle Competition: https://www.kaggle.com/c/prudential-life-insurance-assessment/overview