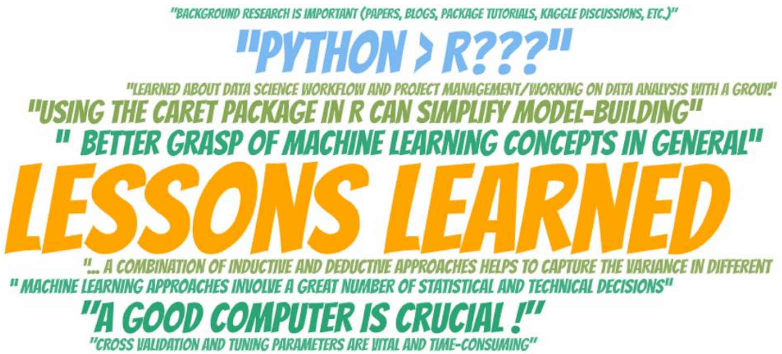
SIOP Machine Learning Competition:
Team Procrustination

Presenters: Feng Guo and Samuel T. McAbee

April 15th, 2021

BGSU®
Bowling Green State University

Building on Past Success



From our 2nd place SIOP (2019) presentation

- Learning is winning
- Aim for simple, stable solution
  - Avoid over-complex models
  - More cross-validation
  - Leverage ensemble models
- Practice good time management
- Effective communication among team members
  - E.g., Google Colab notebook

Project Timeline

- STAGE 01: Planning
- STAGE 02: Data Exploration
- STAGE 03: Feature & Model Selection
- STAGE 04: Model Validation
- STAGE 05: Final Tuning

2020 (TWO WEEKS)

Work Resumes (March 2021)

2021 (3 DAYS)

Stage I
Planning

Brainstorm project ideas
-Validity-diversity trade-off
-Pareto optimal (multi-objective optimization problem)
-Deep Reinforcement Learning?

Stage II

Data exploration, visualization, & preparation (Late Feb 2020)

Data structure and data type

Basic descriptive statistics
-feature intercorrelations
-base-rate for criterions

Missing values

Stage II – Data exploration

Understand the evaluation metric

Final_score = Overall_accuracy – Unfairness

Employees who (are) :

- Top performers? Retained? From Protected groups?

Build your own customized metric

```
# evaluation metrics
def metric(df, y_pred):
```

```
metric(df,y_pred)
```

```
true_top_performers:   0.6413,true_retained:   0.5075,true_top_retained:   0.6347
true_minority_ratio:   0.4950,true_majority_ratio:   0.5023,adverse_impact:   0.9855
accuracy:   60.4542,unfairness:   1.4510,Final_score:   59.0032
```

Stage III & IV

Feature & Model Selection, and Validation

(3/1-3/3, 3/11-3/13, 2020)

Strategy:

- Train multiple models to best predict different components/objectives from the evaluation metric
- Predict “Top performers” “Retained”, “Top performers & Retained” and “Protected groups,” respectively
- Study metric scores to build one simple weighted model to start

Stage III & IV

Features:

- All predictor variables were used to train the model
- Only kept complete rows
  - Special treatment for missing values did not improve our model (e.g., imputation, adding additional features based on the missingness)
  - Additional 36,212 respondents that have turnover data did not further improve our model
- Training data has shape of 6819 (row)x 421 (features)

Stage III & IV

Models:

- Main algorithm: **XGBoost** (eXtreme Gradient Boosting)
  - Improved version of gradient boosting with higher computation efficiency and better capability to deal with over-fitting problems
  - Both XGB regression and classification models were applied

- We also tested deep learning model such as MLP (no improvements)
- Outcome variables selected for separate models:
  - “High_Performer” ;“Overall_Rating” ; “Protected”
- Cross-validation

Summary for 2020

9 submissions for development dataset

- 3rd submission with score of 59.1
- Best submission with score of 60.5

Best model in 2020

- Weighted ensemble model from "High_Performer" , "Overall_Rating", and "Protected" predictions
- **Idea**: First select the top performers, then adjusted by workers from protected groups
- No fine-turning for xgboost model
- Cross validation suggests the solution is not very stable: <u>unfairness score </u>matters most!

Stage III, IV, & V

Model Selection, Validation, and final Tuning

(3/12-3/14 2021)

Picking back up: Two main tasks

- **How to better adjust Unfairness score?**
  - Instead of predicting workers from protected groups, predicting workers who were from protected groups *and* also stayed within organizations
- Fine tune XGBoost model:
  - **Bayesian Hyperparameter Optimization** (e.g., Forest_minimize method)

Final winning model (62.5 for test data)

Final_Zscore = (M1*.5+M2*.3+M3*.2)*.9+M4*.1

Where:

- M1 and M2 are both prediction models for “High_Performer” (The difference is default vs fine-tuned hyper-parameters)
- M3 is the prediction model for “Overall_Rating”
- M4: is the prediction model for “Protected_Group ==1 & Retained ==1”

The employee selection was based on the rankings of the predicted final Z scores, which are the weighted standardized scores by those 4 distinct xgboost models mentioned above

Lessons (re)learned this year

- Ensemble models are working well!
- Simple solutions can be good solutions
- Cross-validation is (really) important
- Great opportunity to apply ML techniques to tackle real problems
- More work is required for model interpretability

Thank you!

For more information on this and related projects, please contact Dr. Sam McAbee (smcabee@bgsu.edu)