Predict Click-Through Rate of Display Ads

Max Wong
Fall 2024

Steps

1. Identify numerical & categorical variables
2. Analyze data (<10% missing for each variable)
3. Prepare data (Imputation)
4. Construct model
5. Predict CTR

Models Tested

- Linear Regression, GAM, Tree, Forest, Boost

Best Model (GAM with 7 variables + smoothing)

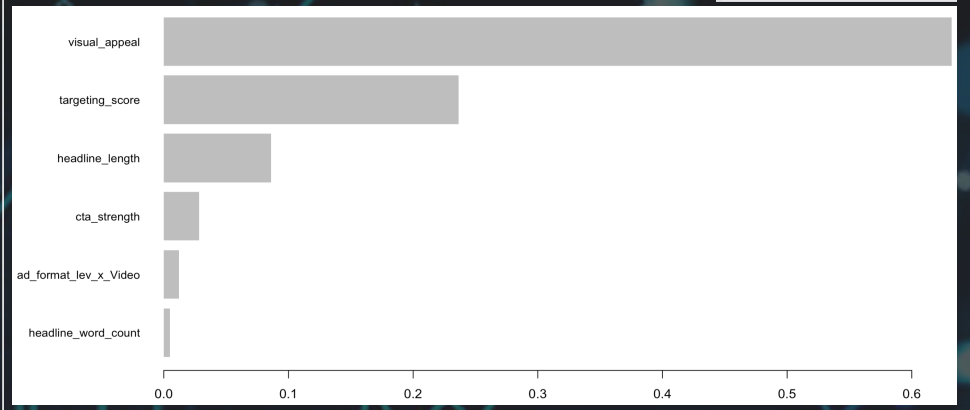
```
model <- gam(CTR ~ s(targeting_score) + s(visual_appeal) + s(headline_length) +
                   factor(time_of_day) + s(brand_familiarity, sp=0.9) +
                   s(cta_strength) + s(body_keyword_density), data = dataClean)
summary(model)
```

- Very small p-value for most variables
- High adjusted R-squared & deviance explained
- Low Generalized Cross-Validation (GCV)

Submitted Model (Xgboost w/ 6 variables)

- 5-fold cross validation
- Find best hyperparameter values

Feature
visual_appeal
targeting_score
headline_length
cta_strength
ad_format_lev_x_Video
headline_word_count



Conclusion & Improvements

- Xgboost model performance on unseen data
  - 10 variables better than 6 variables
  - 6 variables: Overfit on training data
- Consider model choices & hyperparameter adjustments