Otto Group Product Classification Challenge

Stockton Aubrey
Stat 348

Problem:

The Otto Group, a major e-commerce company, faces challenges in analyzing product performance because identical products are often classified differently across their global network. The goal of this challenge is to build a model that predicts product categories based on 93 features for over 200,000 products.

The evaluation metric is multi-class logarithmic loss, which measures how well the model predicts probabilities for the correct category. Better classification will help Otto improve business decisions and customer experiences across their brands, like Crate & Barrel and Otto.de.

Feature Engineering

```
# PREPARE DATA
otto_train <- otto_train %>%
  mutate(target = factor(target)) # Convert target to factor for classification
```

```
otto_recipe <- recipe(target ~ ., data = otto_train) %>%
  step_rm(id) %>% # Remove unnecessary ID column
  step_normalize(all_numeric_predictors()) # Normalize numeric predictors
```

Light GBM Model

```
# LIGHTGBM MODEL
otto_lgbm <- boost_tree(
  trees = 1000,
  tree_depth = 4,
  learn_rate = 0.1
) %>%
  set_engine("lightgbm") %>%
  set_mode("classification")

# CROSS-VALIDATION
set.seed(42)
folds <- vfold_cv(otto_train, v = 5)

cv_results_lgbm <- fit_resamples(
  otto_lgbm,
  otto_recipe,
  resamples = folds,
  metrics = metric_set(mn_log_loss),
  control = control_resamples(save_pred = TRUE)
)
```

Random Forests Model

```
# RANDOM FOREST MODEL
otto_rf <- rand_forest(
  trees = 500,
  mtry = 10,
  min_n = 2
) %>%
  set_engine("ranger") %>%
  set_mode("classification")

cv_results_rf <- fit_resamples(
  otto_rf,
  otto_recipe,
  resamples = folds,
  metrics = metric_set(mn_log_loss),
  control = control_resamples(save_pred = TRUE)
)

collect_metrics(cv_results_rf) # Random Forest CV results
```

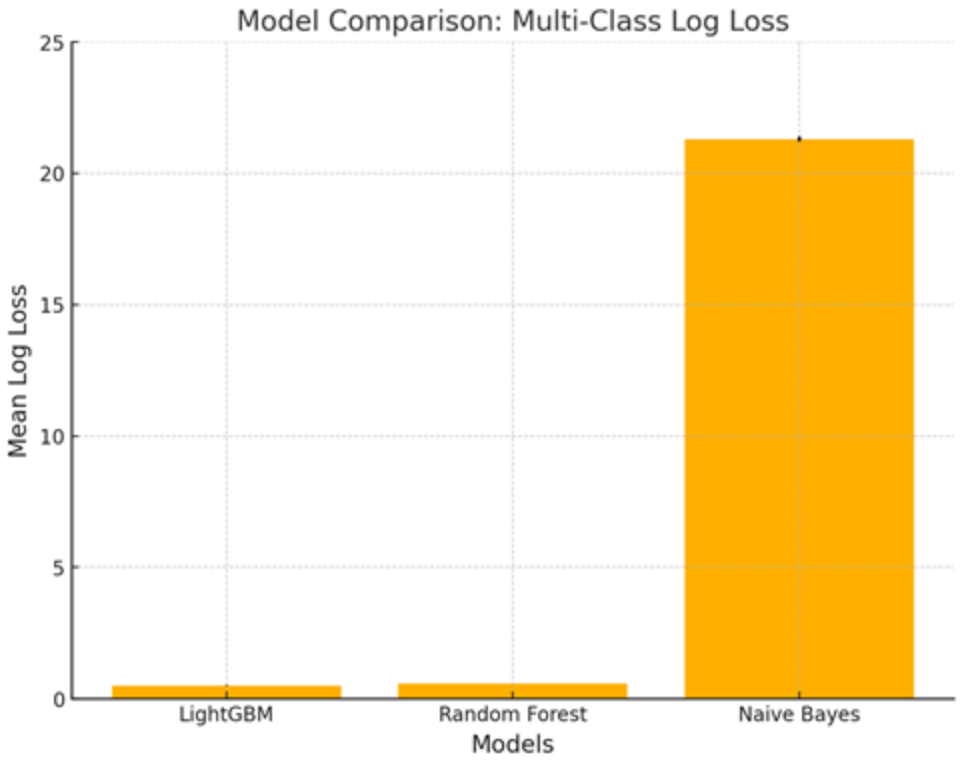
Naive Bayes Model

```
# NAIVE BAYES MODEL
otto_nb <- naive_Bayes(
  Laplace = 0,
  smoothness = 1.5
) %>%
  set_engine("naivebayes") %>%
  set_mode("classification")

cv_results_nb <- fit_resamples(
  otto_nb,
  otto_recipe,
  resamples = folds,
  metrics = metric_set(mn_log_loss),
  control = control_resamples(save_pred = TRUE)
)

collect_metrics(cv_results_nb) # Naive Bayes CV results
```

Model Comparison

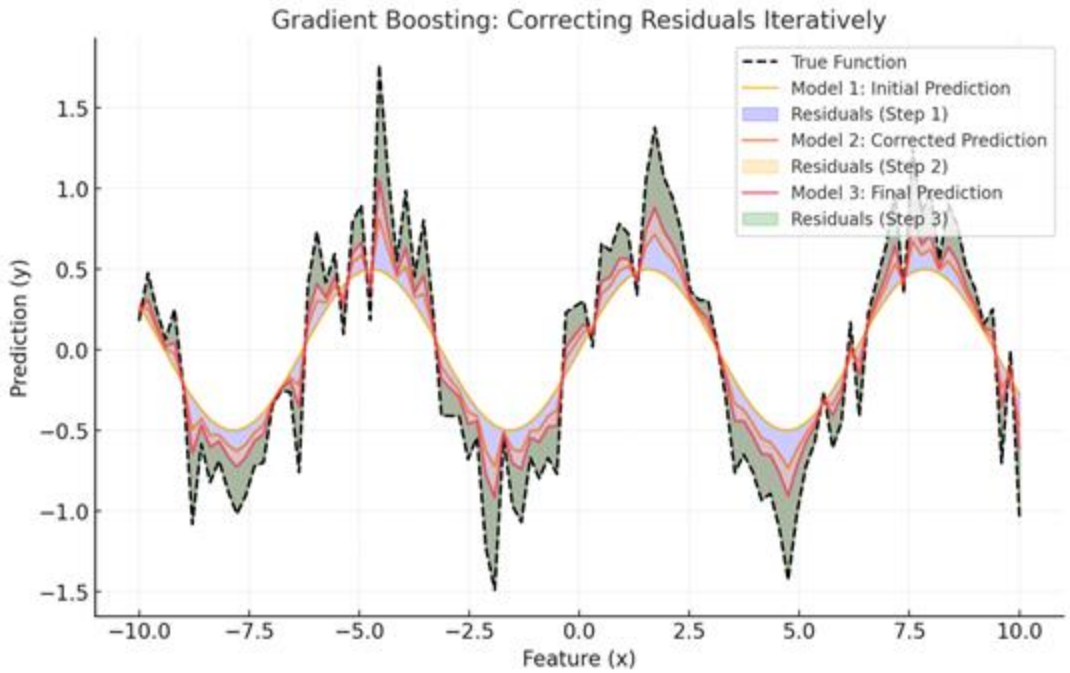



```
# A tibble: 3 x 7
  .metric     .estimator  mean     n std_err .config              Model
  <chr>       <chr>       <dbl> <int>   <dbl> <chr>                <chr>
1 mn_log_loss multiclass  0.499     5 0.00857 Preprocessor1_Model1 LightGBM
2 mn_log_loss multiclass  0.566     5 0.00541 Preprocessor1_Model1 Random Forest
3 mn_log_loss multiclass 21.3       5 0.0930  Preprocessor1_Model1 Naive Bayes
>
```

Why is LightGBM the best?

- Gradient Boosting
- Log Loss Optimization
- Algorithm

Behind Light GBM & Log Loss



Log loss evaluates the entire probability distribution for multi-class classification

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

- N: Total number of samples.
- M: Number of classes (9 in this case).
- y_{ij}: 1 if the sample i belongs to class j, otherwise 0.
- p_{ij}: Predicted probability that sample i belongs to class j.

Score

0.47720