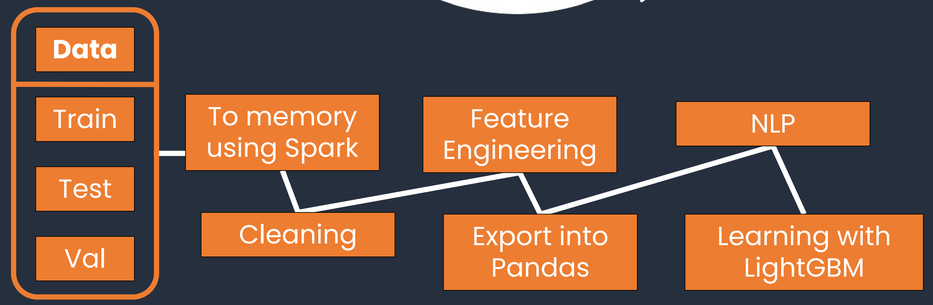
PREDICTING amazon REVIEW HELPFULNESS

ML PIPELINE



DATA CLEANING

- **Clean product features:**
  - product_title: values with length < 4 or NaNs → "Unknown"
  - Replace non-consistent product features with the majority value for each product_id
- **Clean review features:**
  - review_date: only 1 NaN → Drop rows with NAN
  - review_headline: 911 NaNs → Turn into present/not present
  - review_body: remove breakline, replace swapped characters, parse html tags and markers back to formatted unicode text

FEATURE ENGINEERING

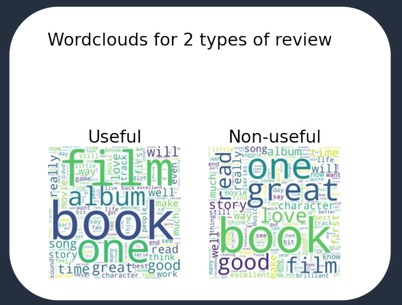
- **Numerical feature construction**
  - Review length
  - Total reviews of product in dataset
- **Categorical feature construction**
  - Year, month and day of the week extracted from review date
  - NLP features

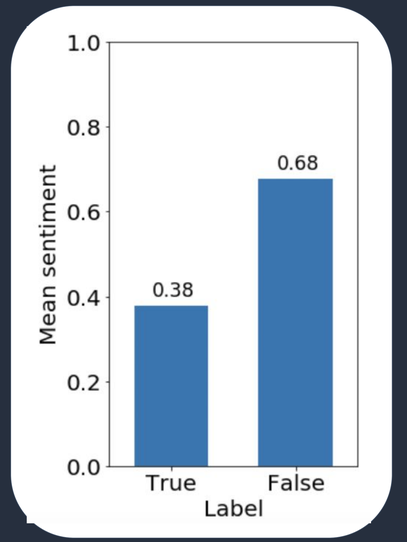
CLASSIFICATION

- **Linear regression:**
  - Baseline model, average performance
- **Fully connected Neural Network**:
  - Moderate performance, handles nonlinearity
- **Light GBM**:
  - Combination of gradient boost and random forest
  - Hyperparameter optimization with Optuna for best result

NLP

- **Review language**
- **Spelling and grammar as text quality indicator**
  - Spelling and grammar mistakes extracted from text
  - Used as ratio of text length
- **Sentiment analysis**
  - All reviews translated with **Google Translate**
  - Sentiment extracted from English text using Flair





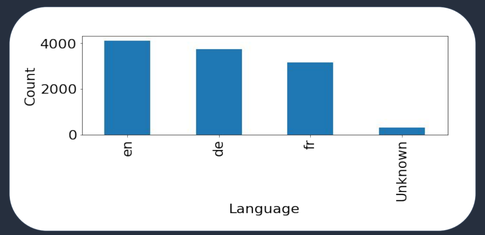
LightGBM

PySpark

pandas

Val Score 0.7742

Test Score 0.7291



Group 6