Fake Review Classification

Anomaly Detection – Challenge 2

Team Pauliguel

Index

- Dataset description
- Undersampling
- Pre-processing
- Models and evaluation
- Tuning
- Kaggle submission

Dataset description

- Data reading problems due to errors in the test reviews file.
- Total of 2908 instances.
- Unbalanced class representation
  - Positive class is 13% of the instances.

Undersampling

- To solve the unbalanced data.
- Random sample of the most common class.
- 784 instances in the training set.

Missing Values

- Uncommon.

- Only one instance.
  - Replace the content of the review by ‘ ‘

Normalization: using the z-score method.

Feature Extraction

- Features from the paper

  Text mining

  - Length review (characters and and words)
  - Bag of words

Personal pronouns	i, we, me, us, my, mine, our and ours.
Associated actions	Went and feel.
Targets	Area, options, price and stay.
Emotion words	Nice, deal, comfort and helpful.

Feature Extraction

- Features from the paper

Behavioral

- Maximum Number of Reviews (MNR)

$$MNR(reviewer_n) = \frac{|reviews(reviewer_n)|}{daysBetween(joinDate(reviewer_n), date(mostRecentReview))}$$

- Percentage of Positive Reviews (PR)

$$PR(reviewer_n) = \frac{|positiveReviews(reviewer_n)|}{|reviews(reviewer_n)|}$$

Feature Extraction

- Features from the paper

Behavioral

- Reviewer Deviation (RD)

$$RD(review_n) = rating(review_n) - rating(hotel(review_n)))$$

- Maximum Content Similarity (MCS)
  - For each Reviewer gather all the Reviews
  - Compute the average of the cosine similarity between each pair of reviews.

Feature Extraction

- Other features
  - Mark fake users (users with at least one fake review)

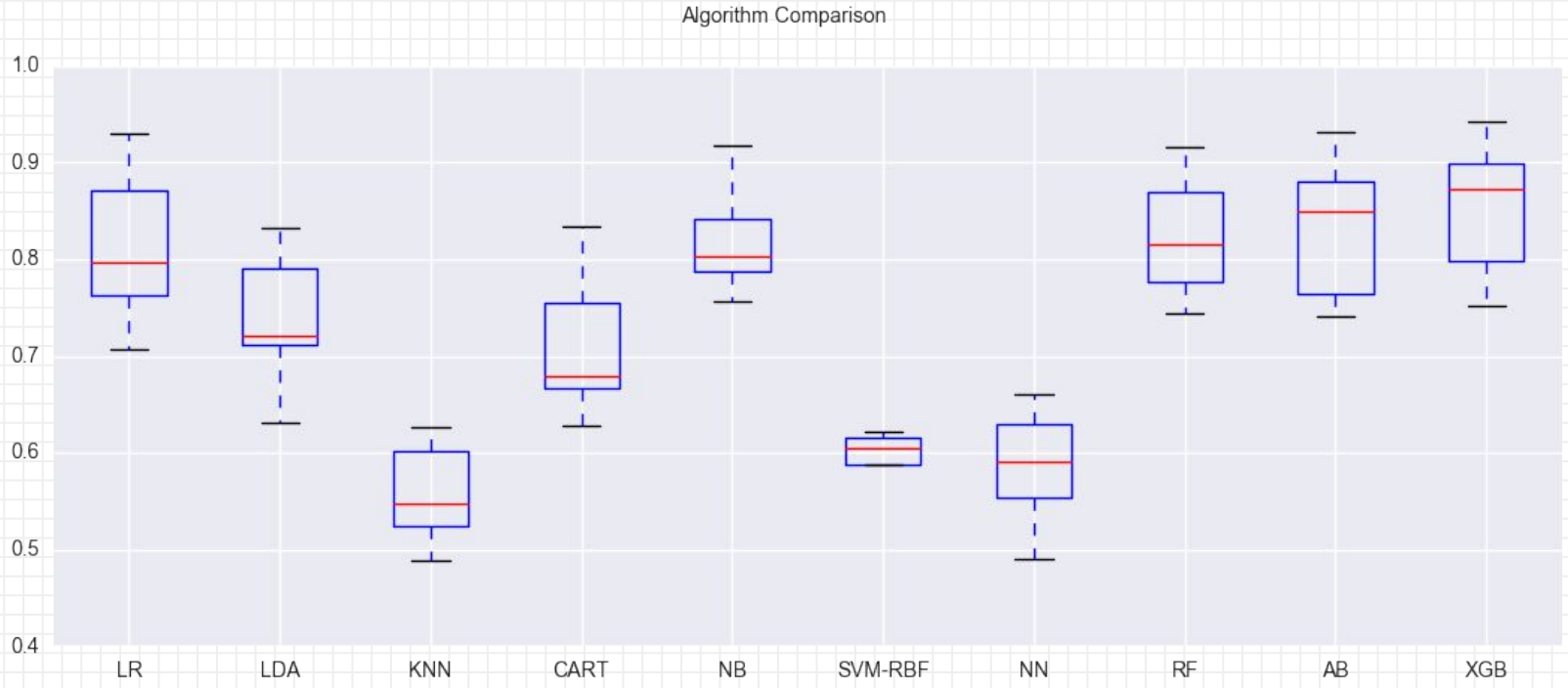
- rating_review - usefulCount_review - coolCount_review - funnyCount_review - friendCount_user - reviewCount_user - firstCount_user - usefulCount_user'	- coolCount_user - funnyCount_user - complimentCount_user - tipCount_user - fanCount_user - rating_hotel - filReviewCount_hotel'

Models and evaluation

Models used:

- Logistic Regression
- Linear Discriminant Analysis
- K-Nearest Neighbors Classifier
- Decision Tree Classifier (CART)
- Naïve Bayes
- Support Vector Machines (with both “poly” and “rgf” kernels)
- Multilayer Perceptron
- Random Forest
- AdaBoost
- Gradient Boosted Decision Trees (XGBoost)

Results with AUC



Tuning

- Random Forests parameters using grid search:
  - “number of estimators”,
  - “criterion”
  - “maximum number of features”.

new score = 0.858400

- XGBoost
  - No increase in the ROC-AUC metric

Kaggle submission

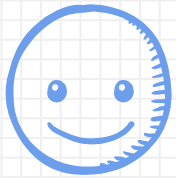
0.78924

Public score - best submission

0.79813

Private score - best submission

We basically quit social life this week to improve our score.
Maybe more luck next time?



THANK YOU!

Any questions?