Random Forest with High Cardinal, Imbalanced Data

Sina Balkhi, Urvi Chauhan, Max Ou, Saboora M. Roshan, and Greg Mori

SFU SIMON FRASER UNIVERSITY

Introduction

Problem Statement

Real-world datasets are usually comprised of numerous categorical features which have high-cardinality for one-hot encoding or other common encoding methods to be effective. Furthermore, these datasets also contain class imbalances.

Literature Review

- Cinema Ensemble Model (CEM) with accuracy of 58.5%
- Movie Investor Assurance System (MIAS) with accuracy of 73%
- Movie success prediction with accuracy of 61% (Random Forest)

Materials and Methods

Data Collection and Preparation

- 2697 movies with 34 features (67% train set and 33% test set)
- The main part of data was collected from TMDB API
- Production budget and revenue were scraped from The Numbers and movie keywords were download from Kaggle
- Encoding categorical features using three methods:

1. CatBoost
$$\hat{x}_i^k = \frac{\sum_{j \neq i}\left(y_i \times (x_j == k)\right) - y_i}{\sum_{j \neq i} x_j == k}$$

2. Leave One Out
$$\hat{x}_i^k = \frac{\sum_{j=0}^{j \leq i}(y_i \times (x_j == k)) - y_i + prior}{\sum_{j=0}^{j \leq i} x_j == k}$$

x_i, y_i: the i – th value and target, k: category

3. Weight of Evidence
$$WOE = \ln\left(\frac{distribution\ of\ good\ credit\ outcomes}{distribution\ of\ bad\ credit\ outcomes}\right)$$

- Target labeling: worldwide gross revenue > budget: label 1; worldwide gross revenue < budget: label 0
- Handling imbalanced data using three Oversampling Methods:
    1. SMOTE
    2. ADASYN
    3. Weight balancing

ADASYN Method

$d = \frac{m_s}{m_l}$; m_s: number of minority class data; m_l: number of majority class data; d: degree of class imbalance

→ $d < d_{th}$ (d_{th}: threshold) → YES →

$G = (m_s - m_l)\ \beta$; G: data needs to be generated; $\beta \in [0,1]$

→ $r_i = \frac{\Delta_i}{K}$; $r_i \in [0,1]$; K: nearest neighbors; Δ_i: number of majority examples in the K nearest neighbors of x_i

→ $\hat{r}_i = \frac{r_i}{\sum_{i=0}^{m_s} r_i}$

→ $g_i = \hat{r}_i \times G$; g_i: data needs to be generated for example x_i

→ For 1 to g_i: choose one minority example randomly (x_{zi}) from the K set for data x_i and generate new minority example (s_i): $s_i = x_i + (x_{zi} - x_i)\ \lambda$; $\lambda \in [0,1]$

SMOTE Method

Compute K neighbors for each minority class data T → Pick one data (n) randomly from neighbors → Features of new minority data (r) = T's features + (T's features – n's features) × rand (0,1)

Classification Methods

- *Decision Tree*

  Entropy: $\sum_{i=1} -p \times \log_2 p_i$ Gain(T,X) = Entropy (T) – Entropy (T,X)

  p_i = Probability of class I, T: target variable, X: Feature to be split on

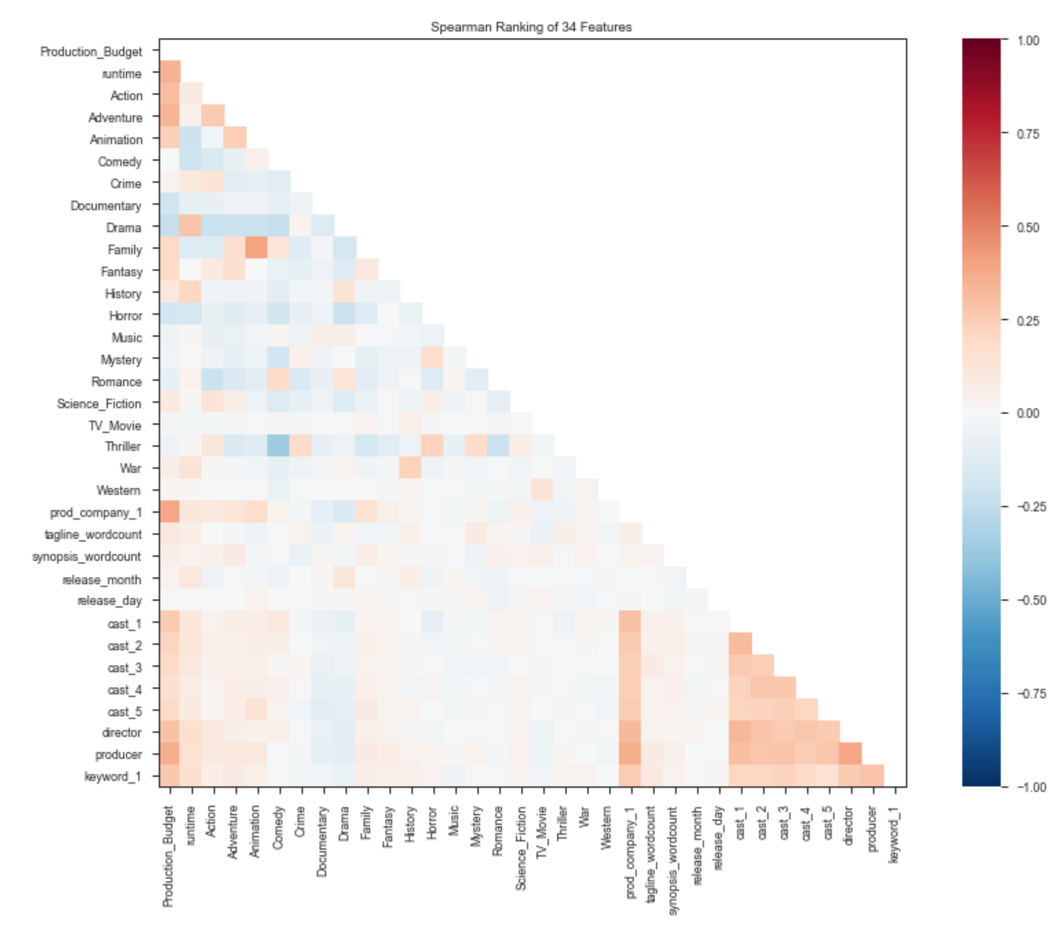
  Entropy (T,X) = The entropy calculated after the data is split on feature X

- *Random Forest*

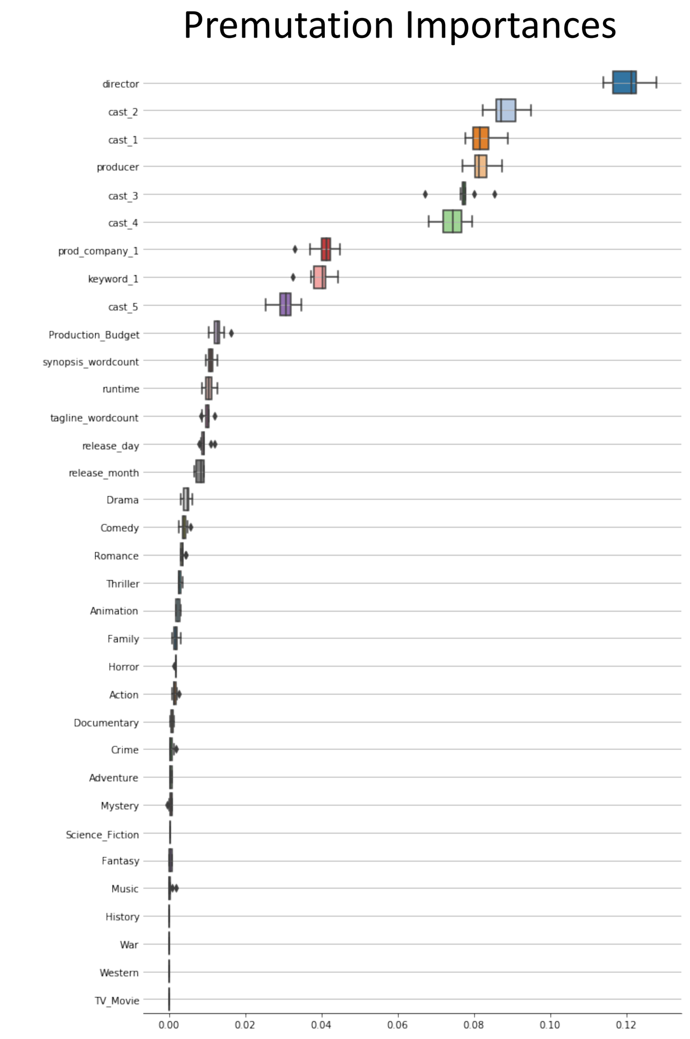
  $\hat{C}_{rf}^B(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B$ $\hat{C}_b$ is prediction of bth random forest tree

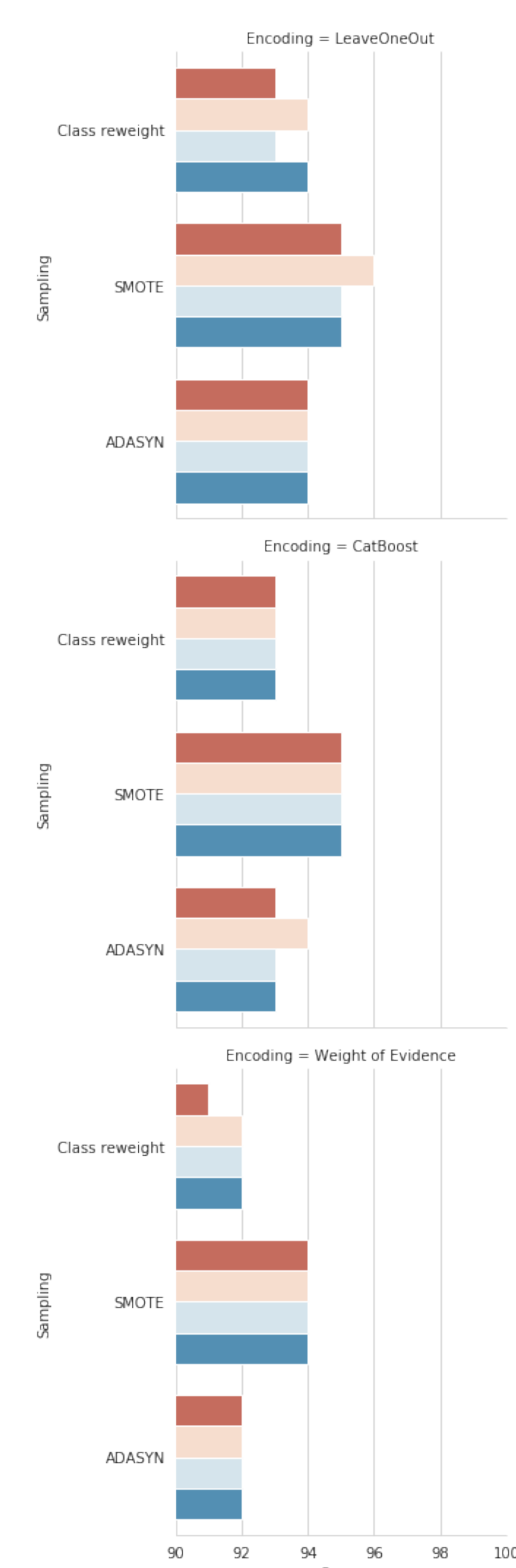
Results and Discussion

Correlation between features

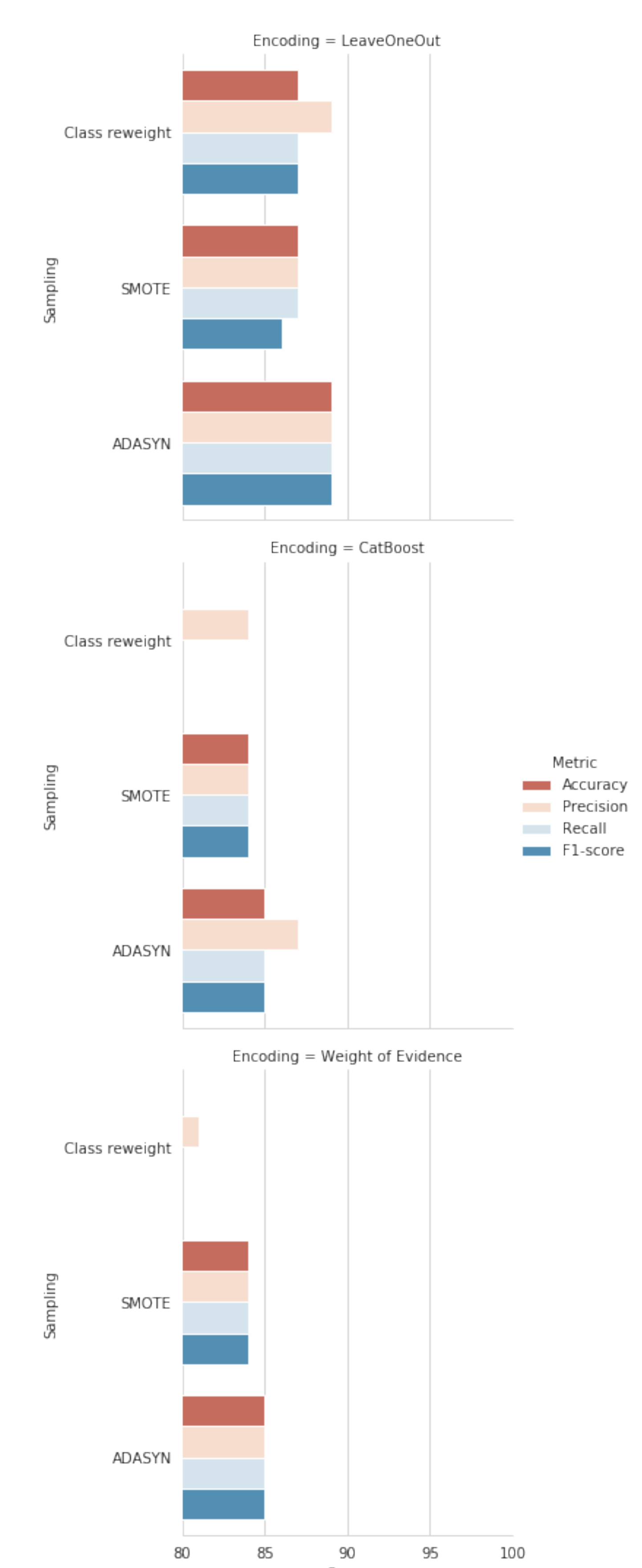


Premutation Importances





Random Forest



Decision Tree

Conclusion and Future Scope

- The best result was in Random Forest method with LeaveOneOut encoding and SMOTE sampling method.
- In the majority case, SMOTE sampling method gave better results that the other approaches
- In future study, we will evaluate our results on different datasets, and also we will work on multiclassification.
- We also suggest that future work should duplicate these experiments on highly correlated data and compare results after removing those correlated features

References

- K. Lee, J. Park, I. Kim et al. "Predicting movie success with machine learning techniques: ways to improve accuracy", Journal of Information Systems Frontiers, Vol. 20, Issue. 3, pp. 577-588, June 2018, [Online]. Available: https://doi.org/10.1007/s10796-016-9689-z
- M. T. Lash and K. Zhao, "Early Predictions of Movie Success: The Who, What, and When of Profitability," Journal of Management Information Systems, Vol. 33, No. 3, pp. 874–903, [Online]. Available: https://doi.org/10.1080/07421222.2016.1243969
- R. Dhir, A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, Dec. 15-17, 2018

Contact Information: sbalkhi@sfu.ca, uchauhan@sfu.ca, maxo@sfu.ca, sabooram@sfu.ca, and mori@cs.sfu.ca