**Random Forest with**

**High Cardinal, Imbalanced Data**

**Sina Balkhi, Urvi Chauhan, Max Ou, Saboora M. Roshan**

**sbalkhi@sfu.ca, uchauhan@sfu.ca, maxo@sfu.ca, sabooram@sfu.ca**

**Abstract**

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 which is a serious problem in classification tasks and may affect the model’s performance. In this project, a comparison study is performed to evaluate different sampling and encoding methods in high cardinal, and imbalanced movie dataset. The movie dataset is collected from various resources and prepared for the binary classification task, in which the success of a movie is being predicted before its release.

**1. Introduction**

In classification tasks, a common problem is imbalanced dataset which causes poor performance in the predictions. Imbalanced dataset problem occurs when instances of a class is much more than the other classes. The class with less instances called “minority class” and the class with more instances called “majority class” [1]. Various fields such as text classification, oil spill detection in radar satellite images, and filtering tasks face imbalanced dataset problem [2]. An example for imbalanced dataset is when 90 % of instances are from majority class and just 10 % are from minority class. In this case, if all predicted instances are majority class, the accuracy of the model would be 90 % which is not a proper metric because none of the minority class are exactly classified [2]. So, imbalanced dataset is a serious problem which affects the model’s performance. There are various methods to handle imbalanced datasets including data level, algorithmic level, and hybrid methods [1, 2]. The well-known approaches of data level are under sampling and oversampling of dataset. Under sampling is to eliminate some instances of majority class to balance both classes. Random under sampling in which the instances removed randomly is the one of the simplest methods in under sampling. There are other sampling techniques such as K-Medoids under-sampling, Condensed Nearest Neighbor, and Neighborhood Cleaning Rule [3, 4].

Under sampling is not a precise method to handle imbalanced dataset, because it may discard useful and important information by removing samples. Random over sampling, Synthetic minority over-sampling technique (SMOTE), and ADASYN are the common methods of oversampling dataset [5]. In addition, random oversampling may cause overfitting by copying some samples [6: why smote2]. Therefore, a new technique was introduced by Chawla et al. (2002), which is called SMOTE. In this oversampling method, artificial samples are generated between each sample from minority class and its k nearest neighbors [7: SMOTE]. Another oversampling method which is introduced by (He et al., 2008) is Adaptive Synthetic (ADASYN) [], and is motivated by SMOTE technique. ADASYN defines higher ratios for minority class samples that are harder to learn compared to the samples that are easier to learn [adasyn]. Another common method for dealing with imbalanced dataset is to assign higher weights for the minority class samples and is called weight balancing [weight].

[6: why smote2] implemented different over sampling methods and introduced SMOTE as the best method for their application. SMOTE was proposed by [ewhy SMOTE 2] as the best method for predicting atrial fibrillation in obese patient. [SMOTE 4] implemented SMOTE for handling class imbalance with the application of osteoporosis patient’s prediction. [2] compared different sampling methods in their survey and introduced SMOTE as a method that performs better in various applications. A new sampling technique is proposed by [smor] to handle ordinal imbalanced dataset which is called Synthetic Minority Ordinal Regression (SMOR). In [smote3], SMOTE and k-means SMOTE are compared for 12 imbalanced datasets from UCI Machine Learning Repository. A comprehensive analysis of SMOTE technique is performed in [smote4] and in this study, SMOTE is suggested for imbalance dataset problem.

The comparative study in this paper, introduces a better understanding of the effects of different sampling and encoding methods in model’s performance when data is highly imbalanced. There are many sampling methods in literature to deal with dataset with imbalanced classes. The most well-known method between all of them is SMOTE. To this end, in this project SMOTE and two other methods along with implementing three encoding algorithms are compared with each other for cardinal, highly movie dataset. Moreover, the state of the art interpretation technique “LIME” is introduced here to evaluate the prediction results.

The rest of report is organized as follows, section 2 briefly explains materials and methods including preparing dataset, sampling methods, encoding techniques, and introducing LIME method. The results and discussion of the comparisons are presented in section 3, while section 4 concludes the project.

**2. Material and Methods**

Data collection and preparation

For this project, 2697 movies with 34 features were used for binary classification. The main part of the data was collected from TMDB API []. This part includes casts, directors, producers, production companies, genres, release date and month, tagline word counts, synopsis word counts. Production budget and revenue were scraped from The Numbers and movie keywords were download from Kaggle []. Target labels are defined as if worldwide gross revenue of a movie is higher than its budget, the label is one and this movie is a success, and if the revenue is lower than budget the label is zero and movie is a failure. For making prediction, dataset is divided to 67% train set and 33% test set. As some of the features such as casts, directors, etc., are categorical dataset, these features are encoded using various methods. In addition, this dataset is cardinal and the common encoding methods like one hot encoding can not be used for encoding because it causes curse dimensionality []. So, three encoding approaches are used here for encoding dataset and their results are compared to each other. These methods are Leave One Out, CatBoost, and Target. Moreover, this dataset is highly imbalanced and 64 % of the data was success and just 36 % was failures. As discussed in introduction section, imbalanced dataset causes poor and inaccurate performance. So, three well-known techniques were used to handle the imbalanced dataset and their results are compared to each other. These methods are SMOTE, ADASYN, and weight balancing. Different encoding and sampling methods are discussed more in the next sections.

**Sampling Methods**

Reweight method

**SMOTE**

In this method, the first step is to compute k nearest neighbors of each minority class sample (). Next, for each , random minority data of its k neighbors is selected and then, the new minority sample is generated by equation ().

|  |  |
| --- | --- |
| + (- ) | (1) |

In this equation, and is a random number [khub1 and SMOTE].

**ADASYN**

This technique defines a ratio for generating new samples. This ratio is based on the majority samples in the neighborhood of the minority class sample. In a simple way, ADASYN assigns a higher ratio for the samples that are difficult to learn and a lower ratio for minority samples that are easy to learn.

The first step of this method is calculating the degree of class imbalance (d), and if then the algorithm will go to the next steps. is a threshold for the maximum

tolerated degree of class imbalance ratio.

|  |  |
| --- | --- |
|  | (2) |

In this equation, and are the number of minority class samples and the number of majority class samples, respectively. Then, the number of minority samples that needs to be generated is calculated by the … equation, in which, is a parameter to specify the desired balance level.

|  |  |
| --- | --- |
| G | (3) |

In the next step, for each sample of minority class, K nearest neighbors are calculated and is computed as follows:

|  |  |
| --- | --- |
|  | (4) |

In this equation, is ratio, is number of majority samples in the K neighbors of . Then is normalized, and so, is a density distribution.

|  |  |
| --- | --- |
|  | (5) |

The number of samples that needs to be generated () for each minority sample is computed in the next step.

|  |  |
| --- | --- |
| = | (6) |

In the last step, for each of each minority sample , a loop from 1 to is done in which, one random minority sample () of neighbors is selected and a new sample is generated by equation ().

|  |  |
| --- | --- |
| = + () | (7) |

In this equation, is a random number and () is the difference vector in the n dimensional space.

**Encoding Methods**

**LIME**

Feature importance is a well-known method that is always used to determine the features’ roles in the predictions and to interpret the model. Another interpretability technique is Local Interpretable Model-agnostic Explanations (LIME). LIME changes a single data and observes the effects of these changes on the output and by this procedure tries to interpret and explain the models. The LIME explanation equation is defined as equation (). By LIME definition, an explanation is a model , in which G is s a class of potentially interpretable models.

|  |  |
| --- | --- |
|  | () |

In this equation, f(x) is the probability (or a binary indicator) that x belongs to a certain class, is a proximity measure between an instance to x, is a measure of how unfaithful g is in approximating f in the locality defined by , and is complexity (as opposed to interpretability) of the explanation g ∈ G [lime].

**Results and Discussion**

As discussed in encoding section, categorical features need to be encoded. Here are the correlation plots using spearman method for three encoding techniques.

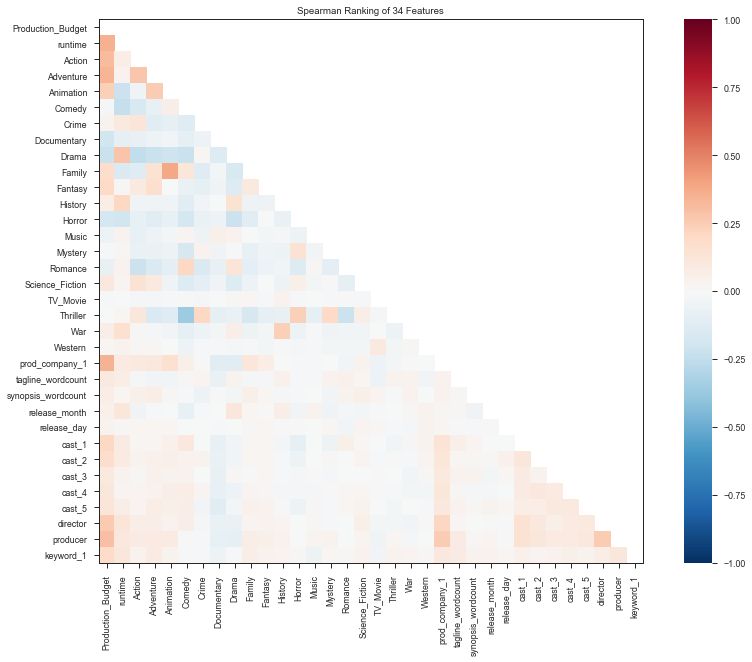


Figure 1. LOO correlation plot

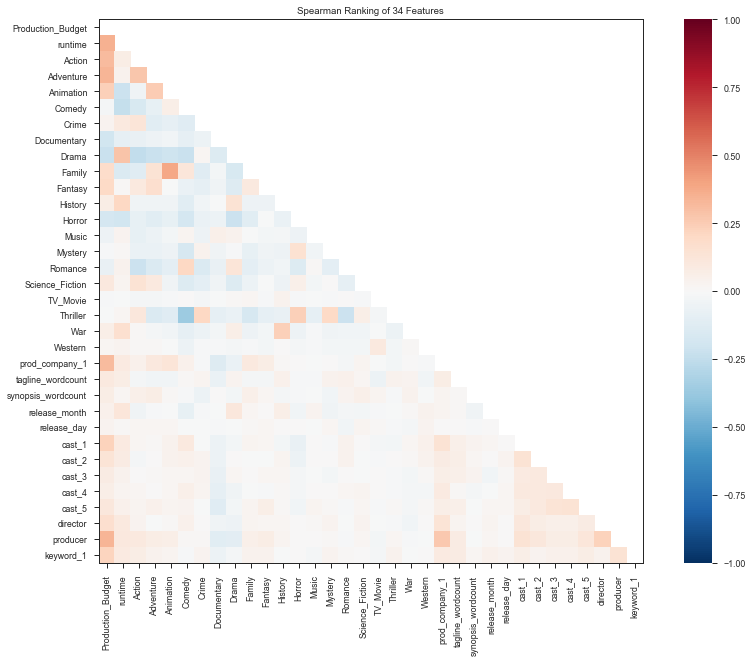


Figure 2. CatBoost correlation plot

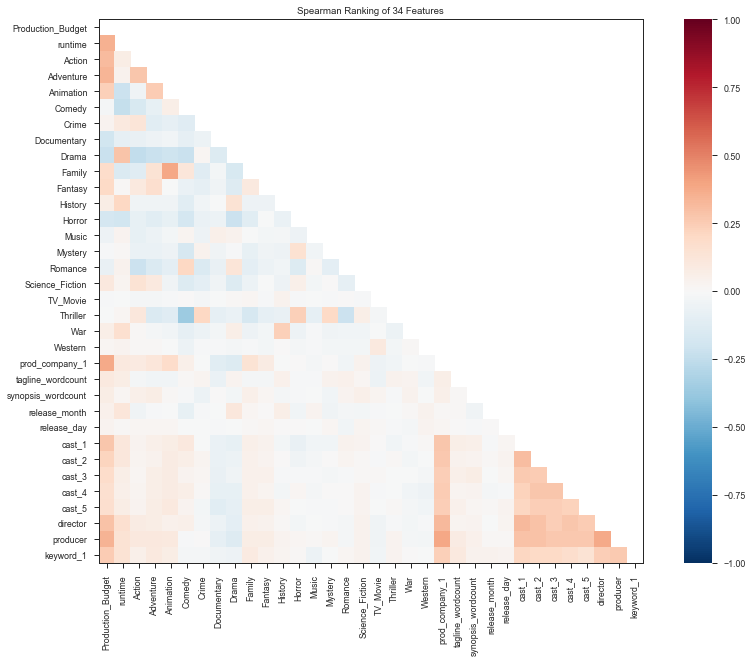


Figure 3. WOE correlation plot

RF and DT graphs or tables

The best result was in Random Forest method with LeaveOneOut encoding and

SMOTE sampling method.

LIME results and explanation

Conclusion and Future Scope

In this study,

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

Contributions

All authors contributed equally to this project.

References