Predict whether a video game will be a blockbuster or not

Niccolò Morabito, Md Jamiur Rahman Rifat, Universitat Politècnica de Catalunya - BarcelonaTech

Video games are a popular medium in the entertainment business that bears a significant economic impact. In our project, we want to predict whether a video game will be a blockbuster or not. To address this, we are using a dataset collected by Dr Joe Cox from 2004 to 2010, containing 1212 instances with 36 features.

PREPROCESSING

- **EDA**. The exploratory phase showed:
  - that most of the numerical features have more than 25% of their values equal to 0;
  - which columns can be removed because they do not contain additional information;
  - the distribution of the numerical features and the composition of the categorical features.
- **Imputation and scaling**:
  - the missing values are imputed with KNN Imputer;
  - all the features are scaled with a StandardScaler.
- **Visualization and deletion of outliers**.
- **Encoding categorical features** with one-hot encoding.
- **Binning target variable** according to statistical intuitions and domain knowledge in two classes.
- **Feature selection and correlation.**

MODEL SELECTION AND HP TUNING

5 machine learning models were tested with 4 experimental setup, tuning HP for each of them with cross validation:

- **Setup1**: with all features
- **Setup 2**: Removing correlated features
- **Setup 3**: excluding non-important features (score < 6)
- **Setup 4**: excluding non-important features (score < 1)

The machine learning models are **Decision Tree**, **Random Forest**, **Support vector machine, Multi Layer Perceptron Network**, **XGBoost Classifier**.

Model	DT	RF	MLPN	SVM	XGB
Setup 1	0.828	0.850	0.828	0.823	0.857
Setup 2	0.827	0.851	0.848	0.849	0.853
Setup 3	0.822	0.846	0.826	0.835	0.852
Setup 4	0.817	0.835	0.822	0.827	0.832

EVALUATION

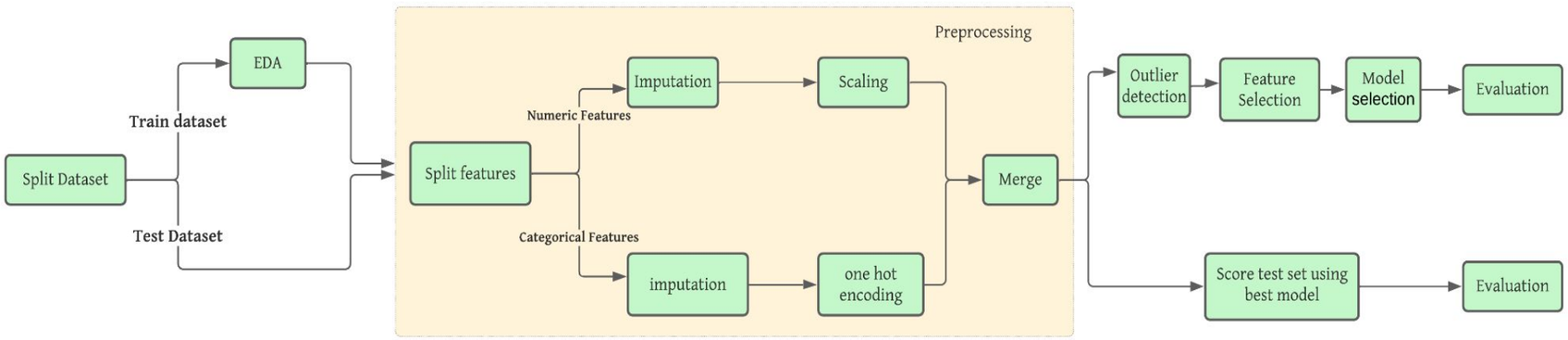
The best obtained model which is “XGB Classifier” has been finally tested on unseen data to get the final evaluation. The values of the most important metrics are summarized in the following table:

	Precision	Recall	F1-score	Support
no	0.84	0.94	0.88	225
yes	0.73	0.47	0.57	78

The negative class has good results since reaches an f1-score of 0.88. Even though the results could be improved (especially the precision, which is 84%), we can definitely say that the problem relies on the positive class, where the results are worse (the recall cannot exceed 50% because of the high number of false negatives).

The number of positive instances in the training set was insufficient and it made the model tend to overfit the negative instances. The unbalance between classes depends especially on the binning we decided to apply (making the positive class include only the video games with a score greater than 80/100), which actually increased the difficulty of the task reducing the number of instances for the positive class and blurring the edge between the two classes (negative class not only included very bad video games with a score close to 0 but also acceptable or valuable games with high scores).

PIPELINE



CONCLUSIONS

The class imbalance is causing the results to be better for the negative class and unsatisfying for the positive one. In order to improve the performance of the model, we should incorporate some methods which take care of the class imbalance problem such as custom weight values and oversampling (undersampling, with the data at our disposal, is not possible because the number of instances is already limited). Moreover, complex neural network models are found to be useful so we would like to use a complex neural network model with more layers and nodes. The effort should focus on improving the capability of the model to predict the positive class and to reduce the number of false negatives.