CSCI6917: Guided Research Methods

Summer 2023

Stephen H. Kaisler, D.Sc. (GWU)
Jamaladdin Hasanov, Ph.D. (ADA)

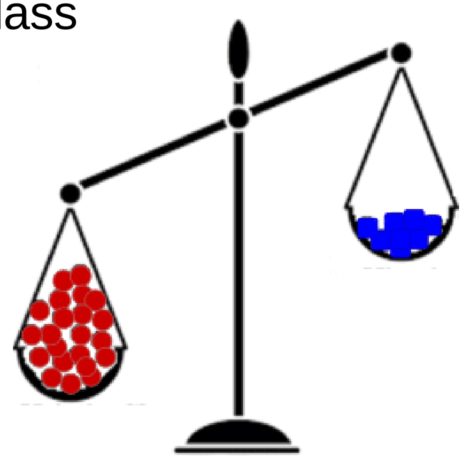
Exploring Class Imbalance Solutions: Investigating the Effectiveness of Data Balancing Techniques on Model Performance

Student: **Shamil Shukurov**

Project Objective

What is class imbalance problem?

- Occurs when one class has significantly fewer instances than the other class
- Imbalanced datasets can bias models towards the majority



What is balancing?

- Techniques used to address class imbalance in datasets and equalize the representation of different classes.
- Balancing aims to prevent model bias towards the majority

Purpose of this project is to answer to following questions:

- How different balancing techniques affects model performance?
- When we need to balance our data?
- Do we even need balancing?
- How balancing affects generalization capability of model?
- Which balancing techniques are effective?

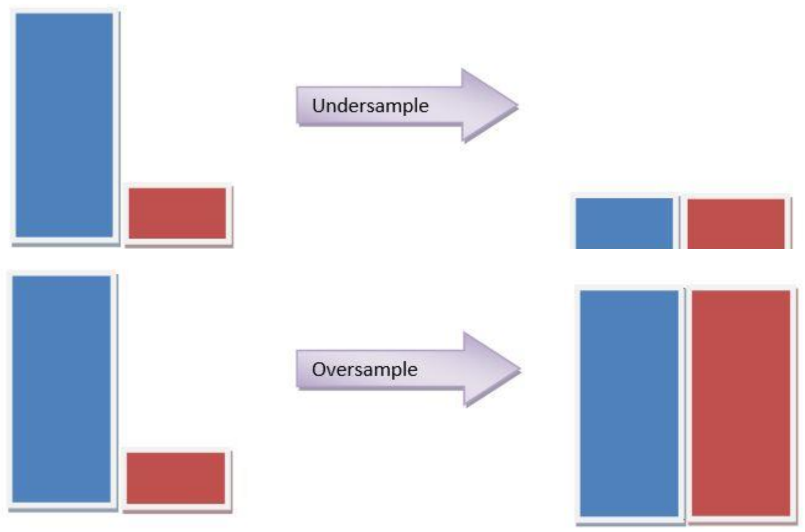
In total, 96 models were built with 5 different balancing techniques and 1 baseline model.

Project Objective

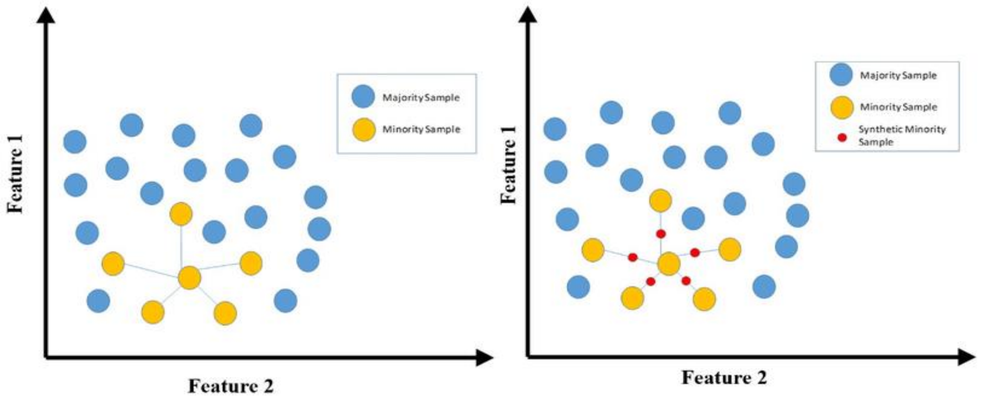
- **How is it done today?**
  - Basic data balancing techniques (upsampling, downsampling, SMOTE) are commonly used to address class imbalance.
  - However, the effectiveness of these techniques is often based on trial and error rather than a comprehensive evaluation.
- **What is new?**
  - Insights on technique effectiveness, class imbalance percentage impact, and overfitting risks.
- **Who will benefit from the work? Why?**
  - Researchers and practitioners dealing with imbalanced datasets.
  - Practical guidance for selecting appropriate balancing techniques and utilizing LightGBM's built-in balancing.
  - Informed decisions to enhance model performance in real-world applications efficiently.
- **What is innovative?**
  - Comprehensive Evaluation with Real and Synthetic Datasets
  - Focus on Class Imbalance Percent
  - Algorithm Built-in Balancing Advantages tested

Technical Approach

- Collection of Data
- Data Preprocessing
- Building Baseline Model
- Applying Balancing techniques to the data
    - Upsampling
    - Downsampling
    - SMOTE
    - BalancedBaggingClassifier
    - Algorithm built-in balancing
- Analyze model results:
    - F1-Score
    - AUC



Downsample and Upsample



SMOTE

Technical Approach : Data

- During the research 13 different imbalanced datasets collected

- Different minority class percentages will show how balancing techniques performs in different degrees of imbalance

- Applied preprocessing steps:
  - Redefining target if needed
  - Encoding the categorical variables
  - Correcting type inconsistencies
  - Imputing missing values if needed

Dataset_Name	Row_Count	Minority_Class_Percent	Target_Column
Fraud	284807	0.172	Class
Fraud2	250000	0.5	is_fraud
Wine	6497	3.76	target
~~Letter-a~~	~~20000~~	~~3.94~~	~~letter_a~~
Abalone	1477	4.52	target
Pendigits	10992	9.59	is_9
Sick_euthyroid	2194	10.05	target
Covertype	581012	14.77	target
~~Letter-vowel~~	~~20000~~	~~19.39~~	~~is_vowel~~
Contraceptive	1743	22.6	target
~~Splice-junction~~	~~3186~~	~~24.07~~	~~target~~
Adult	32561	24.08	target
Churn	7043	26.53	target

Technical Approach : Data

- Problems in collected datasets:
  - Easier classification problems
  - Size of datasets
  - Nature of the data affects metrics

- Additional synthetic data generated:
  - Class distribution: 30%/70%

- By taking samples from this dataset, 5 more datasets generated

```
X, y =
make_classification(
    n_samples=200000,
    n_features=10,
    n_informative=7,
    n_classes=2,
    random_state=42,
    weights=[0.7,0.3],
    hypercube=False,
    class_sep=0.01,
    flip_y=0.15
)
```

Dataset	Minority_Class
Synthetic	0.3
Synthetic_20	0.2
Synthetic_10	0.1
Synthetic-5	0.05
Synthetic-1	0.01
Synthetic-0.5	0.005

Technical Approach : Model

LightGBM is a gradient boosting framework that utilizes a tree-based learning algorithm to construct powerful ensemble models. It offers several advantages that make it well-suited for imbalanced classification tasks.

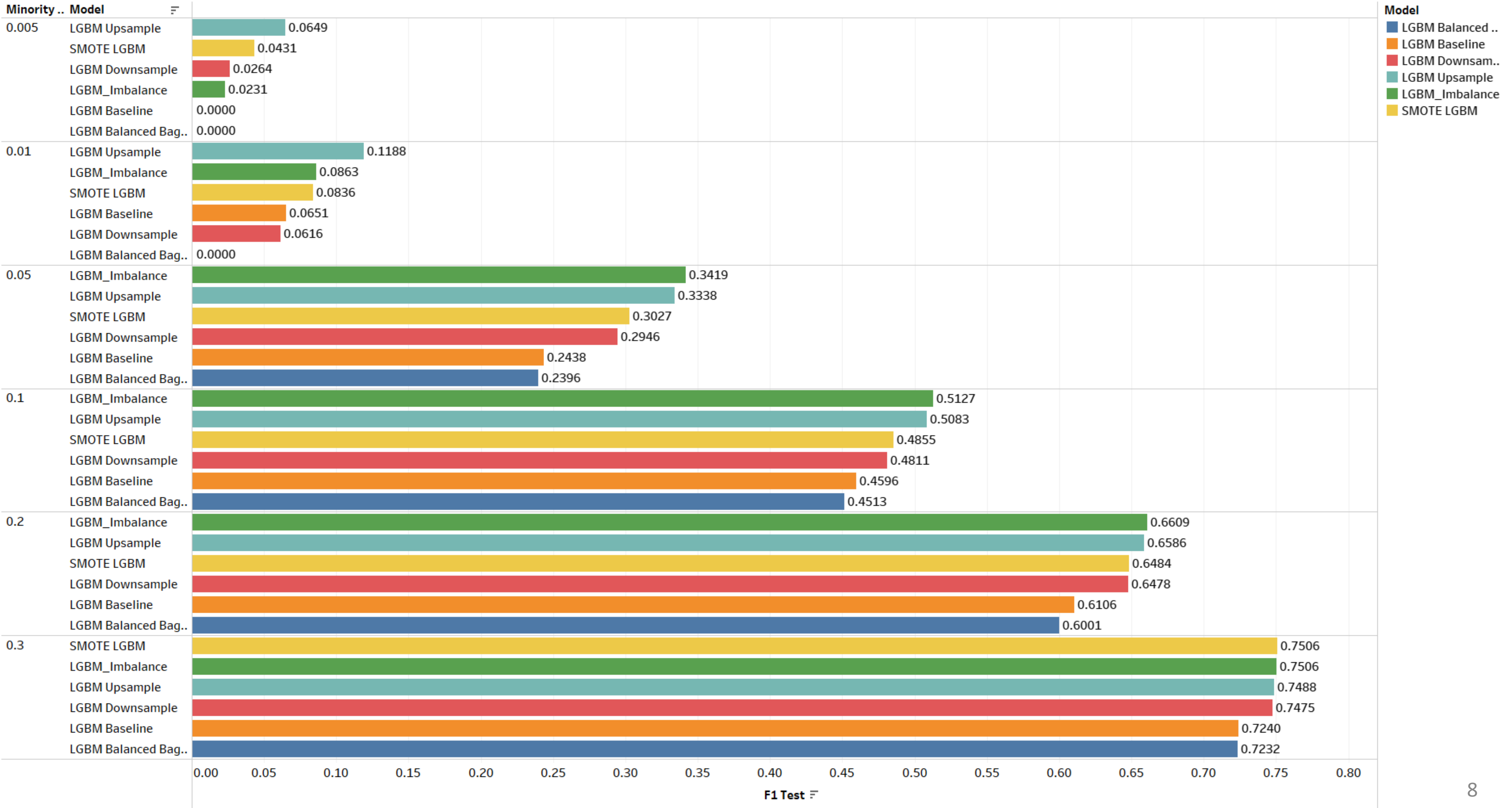
Why LightGBM?

- Built-in imbalance technique
- Doesn’t need a lot of preprocessing
- Popular in practice

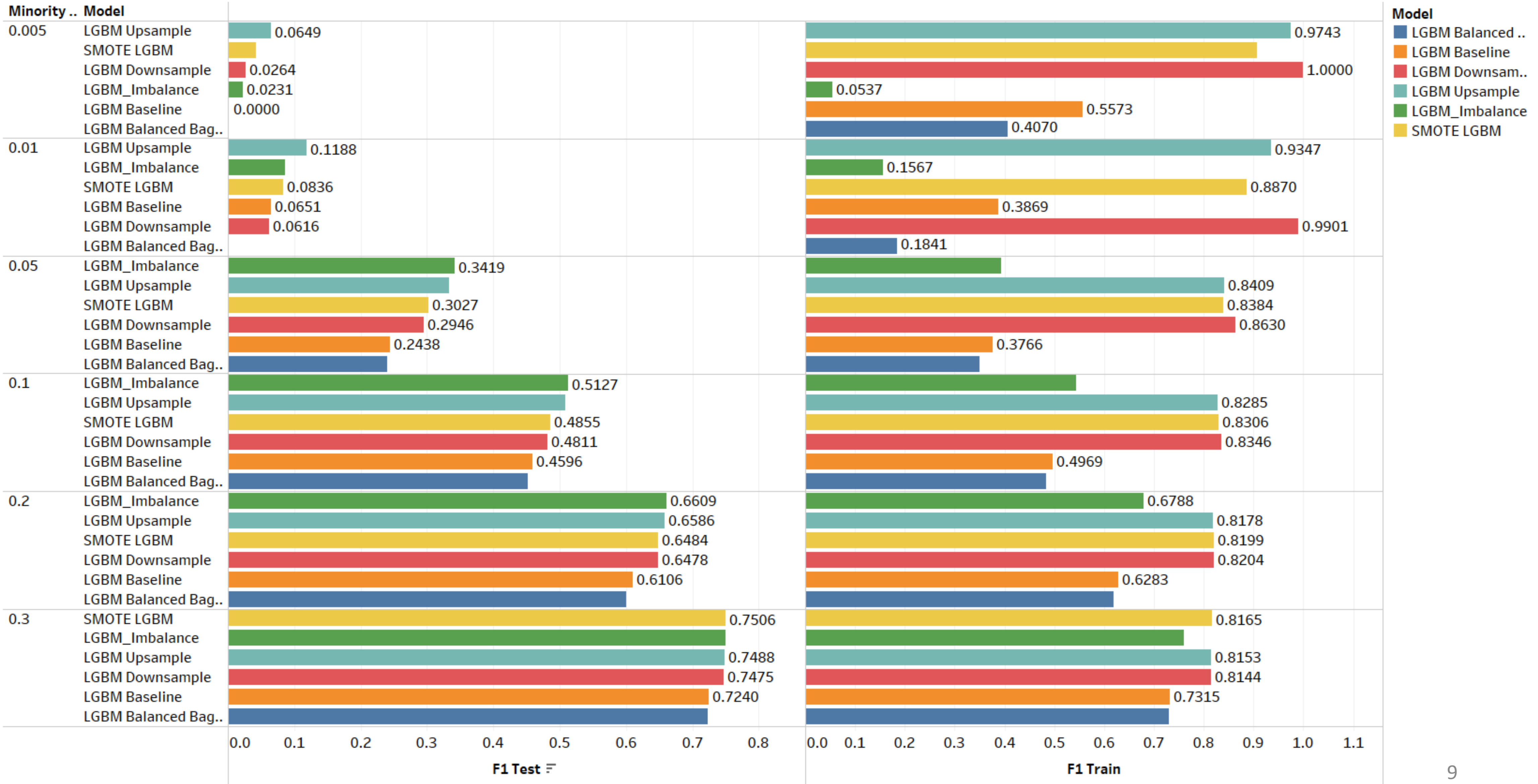
Models:

- LGBM Baseline
- LGBM Upsample
- LGBM Downsample
- SMOTE LGBM
- LGBM Balanced Bagging
- LGBM_Imbalance

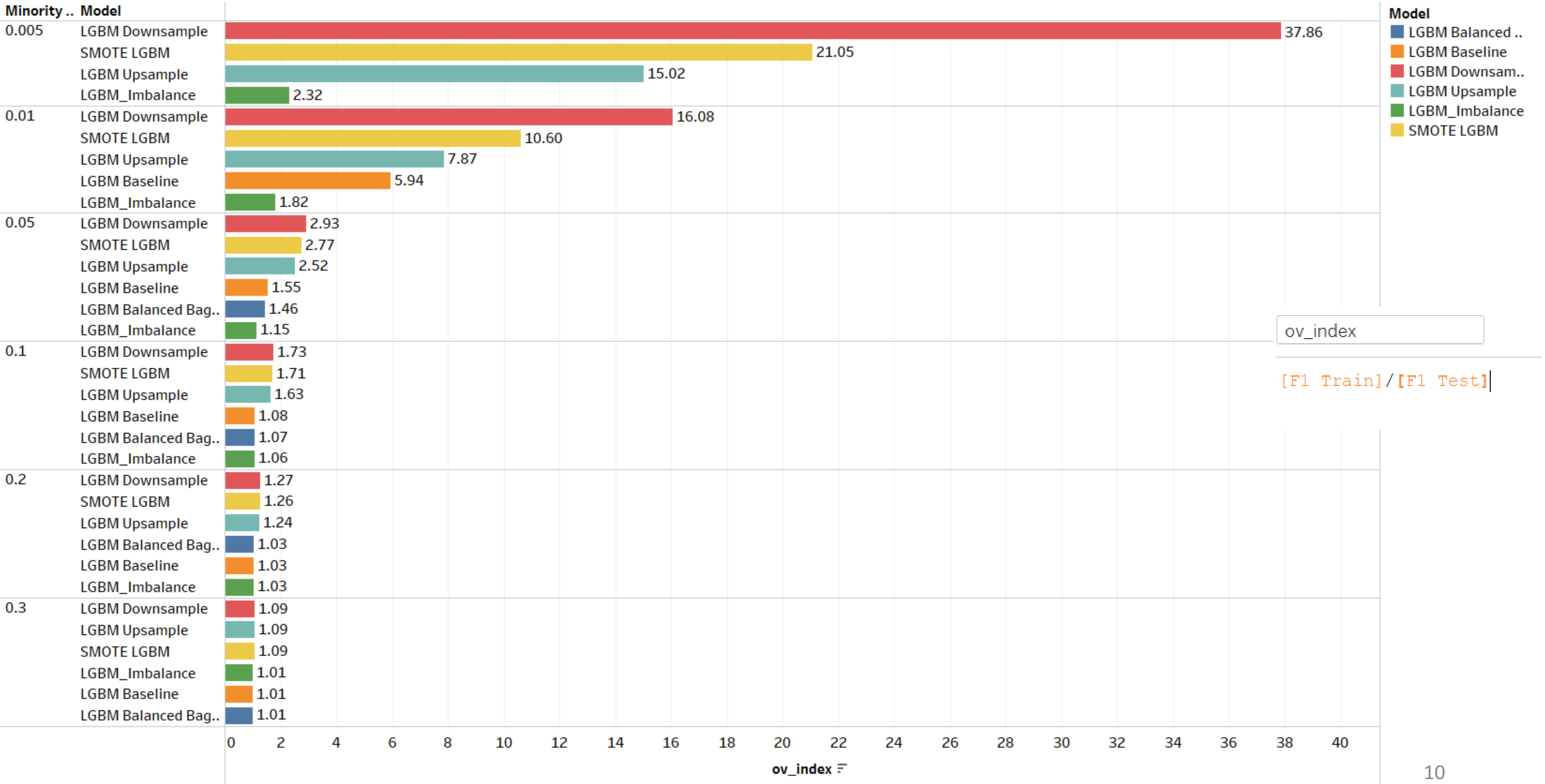
Results : F1-Scores on test data



Results : Overfitting

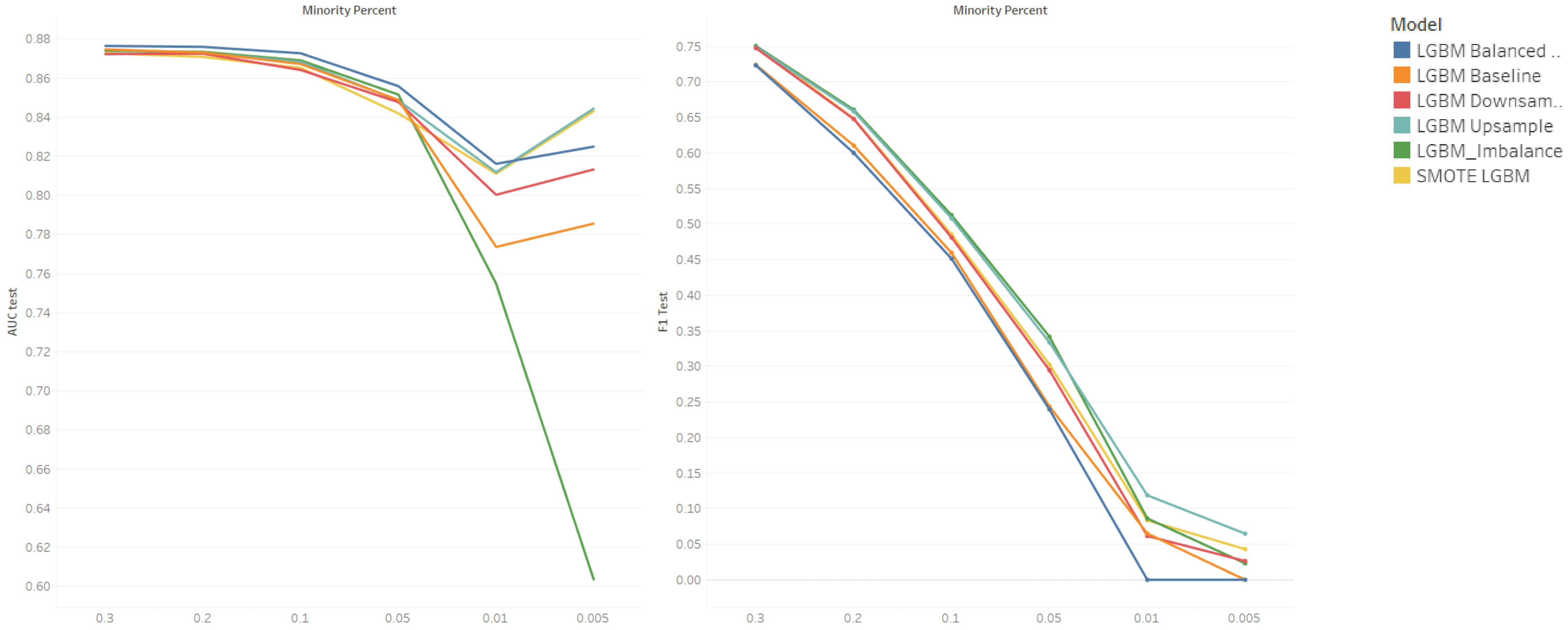


Results : Overfitting



Results

- From minority class percent 30% to 10%, the performances of balancing techniques and the baseline model are very close.



Conclusion

- When working with imbalanced dataset always build baseline model without any balancing technique
- We observed that balancing techniques are not very effective with class imbalance percent > 10%
- Some balancing techniques, especially downsampling, SMOTE and upsampling can cause overfitting
- LGBM's built-in balancing generalizes data better than above techniques

Future Work

- Extending research with:
  - More broad datasets
  - More balancing techniques
  - Tests on different algorithms
- Calibrating models trained on balanced data
- Explaining mathematical reasons behind conclusions

Any Questions?