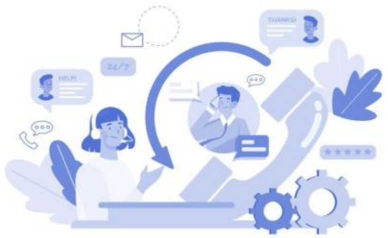
ML Model for Bank Marketing Campaign

CS 613 Final Project

Jenny Boroda, Alec Peterson, Yifan Wang

Problem



- Predict success of marketing calls for selling bank long-term deposits (or other banking/financial products)
- Develop classification machine learning models to inform marketing employees and:
  - Support client selection decisions
  - Reduce time, costs, # of calls
  - Reduce client intrusiveness, improve overall satisfaction

Related Work

- The dataset was collected as part of the direct marketing campaigns of a Portuguese banking institution. The campaign was conducted via phone and was directed to sign up clients to the bank's term deposits.
- The data was collected from May 2008 to June 2013, in total of 52944 phone contacts.
- The introductory paper was published in 2014 by Sergio Moro, P. Cortez, P. Rita in Decision Support Systems.
  - Citation: *Moro, S., Cortez, P. & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems. 62, 22-31*

Related Work

- Paper authors reduce number of features from 150 to 22 through:
  - Business domain knowledge (manual)
  - Optimizing receiver operating characteristic (ROC) curve area (AUC) (automatic)
- Evaluated classification algorithms based on AUC:
  - logistic regression (LR)
  - decision tree (DT)
  - support vector machine (SVM)
  - neural network (NN) - ***best***

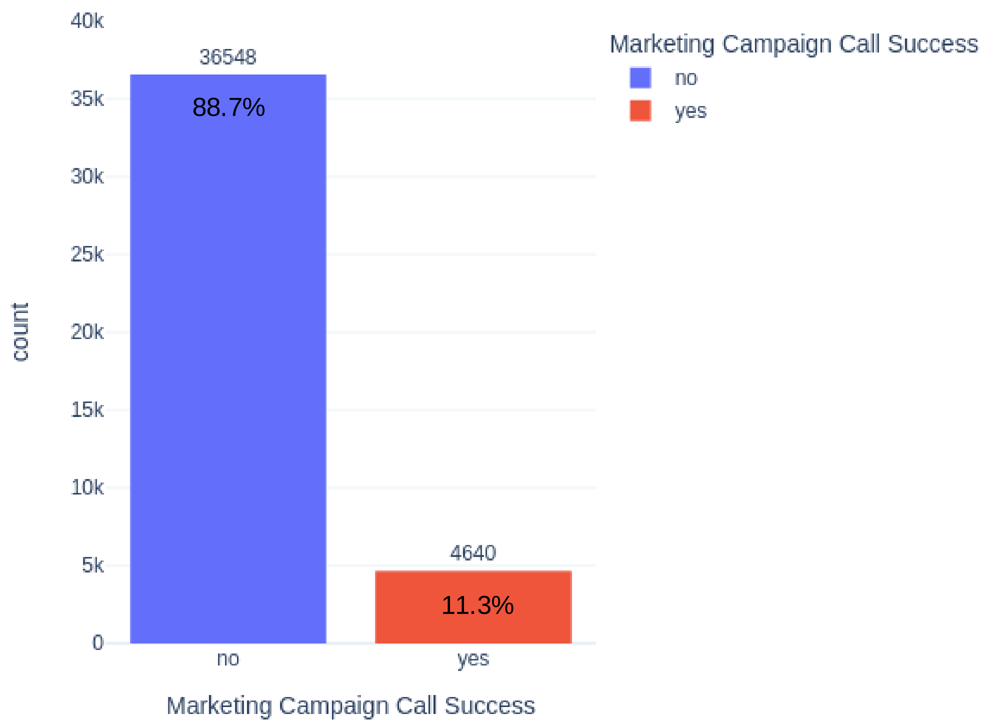
Comparison of DM models for the modeling phase (**bold** denotes best value)

Metric	LR	DT	SVM ($\tilde{\gamma} = 2^{-7.8}, C = 3$)	NN ($\tilde{H} = 6, N_r = 7$)
AUC	0.900	0.833	0.891	**0.929***

- Other groups have evaluated more sophisticated neural networks (Zhu et al., 2018) and tree-based ensembles (Yoon et al., 2017) for similar classification tasks on this dataset and saw improved performance

Class Labels

- Data from Moro, Cortez, and Rita available at [UCI ML Repository]
- Direct marketing campaign success for Portuguese bank
- ~41,000 samples
- Class labels imbalanced:
    - ~89% no, ~11% yes



Features

Numeric

Categorical

#	Feature	Attribute Grouping	Description
1	age	Bank Client Data	Age of client contacted
2	job		Type of job (e.g. "admin.", "unemployed", "management", "student", "blue-collar", "self-employed", etc.)
3	marital		Marital status (married, divorced/widowed, single)
4	education		Extent of education (primary, secondary, tertiary, unknown)
5	default		If client has credit in default (yes, no)
6	housing		If client has a housing loan (yes, no)
7	loan		If client has a personal loan (yes, no)
8	contact	Current Campaign	Contact communication type (telephone, cellular, unknown)
9	day_of_week		Last contact day of the week
10	month		Last contact month of the year
11	duration		Last contact duration, in seconds
12	campaign	Previous Campaigns	Number of contacts performed during this campaign for the client (including last contact)
13	pdays		Number of days that have passed since client was last contacted from a previous campaign (-1 label for not previously contacted)
14	previous		Number of contacts prior to this campaign for the client
15	poutcome		Outcome of previous marketing campaign (success, failure, unknown, other)
16	emp.var.rate	Social and Economic Indicators	Employment variation rate, with a quarterly frequency
17	cons.price.idx		Monthly average consumer price index
18	cons.conf.idx		Monthly average consumer confidence index
19	eurobor3m		Daily three moth Euribor rate
20	nr_employed		Quarterly average of the total number of employed citizens

Pre-Processing

1. Maintain class label proportion (11% for "success") across shuffled-then-split Training Set (2/3) and Validation Set (1/3)
2. Resampled training set so proportion of class labels was even
3. Categorized numerical variables
4. One-hot encoded variables with multiple categories, giving a total of 81 features
5. Optimized algorithm hyperparameters using 5-fold cross-validation and selecting for highest area under ROC curve (AUC) averaged across folds

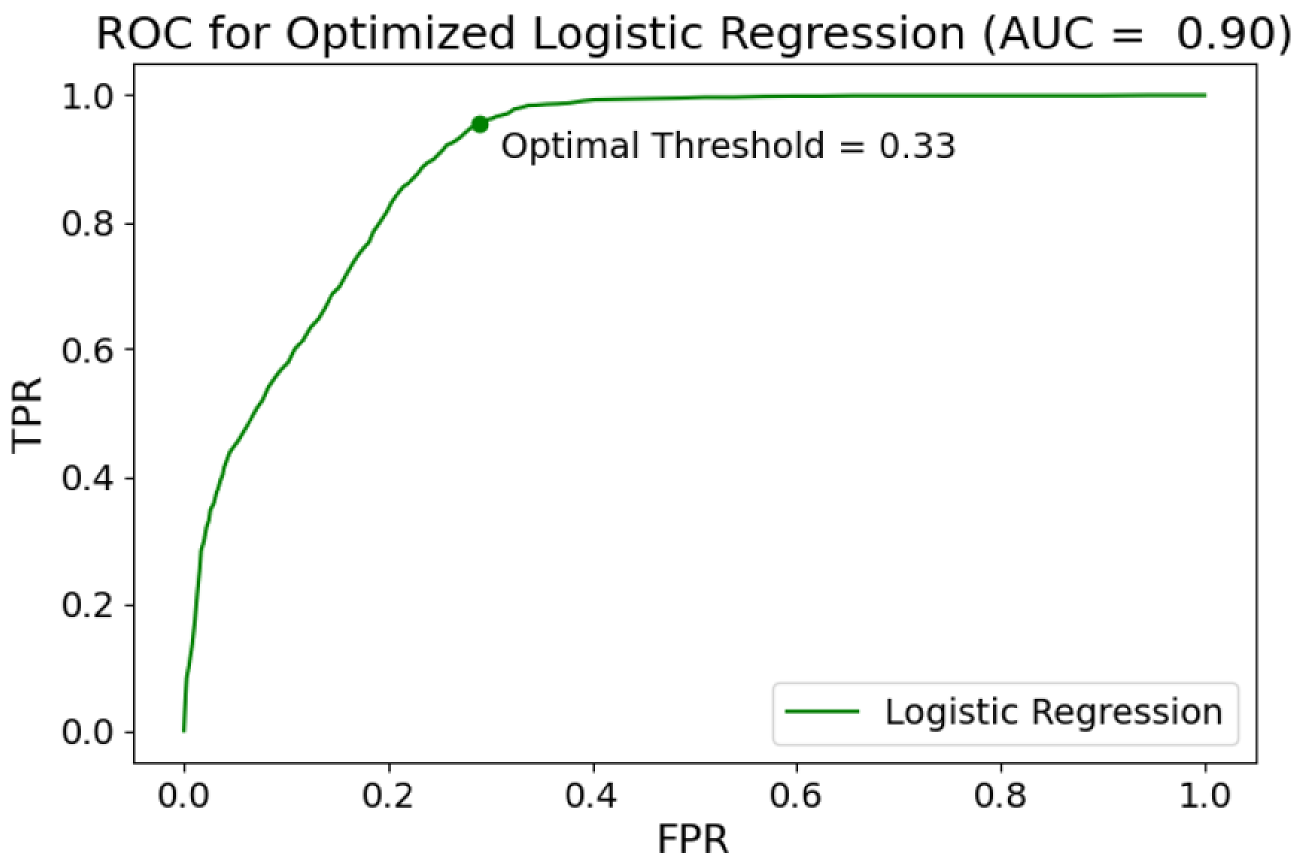
Categorized Numerical Features

#	Feature	Attribute Grouping	Categories
1	age	Bank Client Data	Split into categories: 17-30, 30-40, 40-60, 60-80, 80-100 years
11	duration	Current Campaign	Converted seconds to minutes, then split into categories: 0-1, 1-2, 2-5, >= 5 minutes
12	campaign	Previous Campaigns	Split into categories: 0, 1, 2, 3-5, >5 contacts
13	pdays		Split into categories: -1-0, 0-180, 180-999 days since last contact
14	previous		Split into categories: 0, 1, 2, 3-5, > 5 prior contact
16	emp.var.rate	Social and Economic Indicators	Split into binary category: 1 if >= training set mean, else 0
17	cons.price.idx		
18	cons.conf.idx		
19	eurobor3m		
20	nr_employed		

Logistic Regression

Metric	Value
AUC	0.90
Precision	0.30
Recall	0.96
F1	0.45
Accuracy	0.74

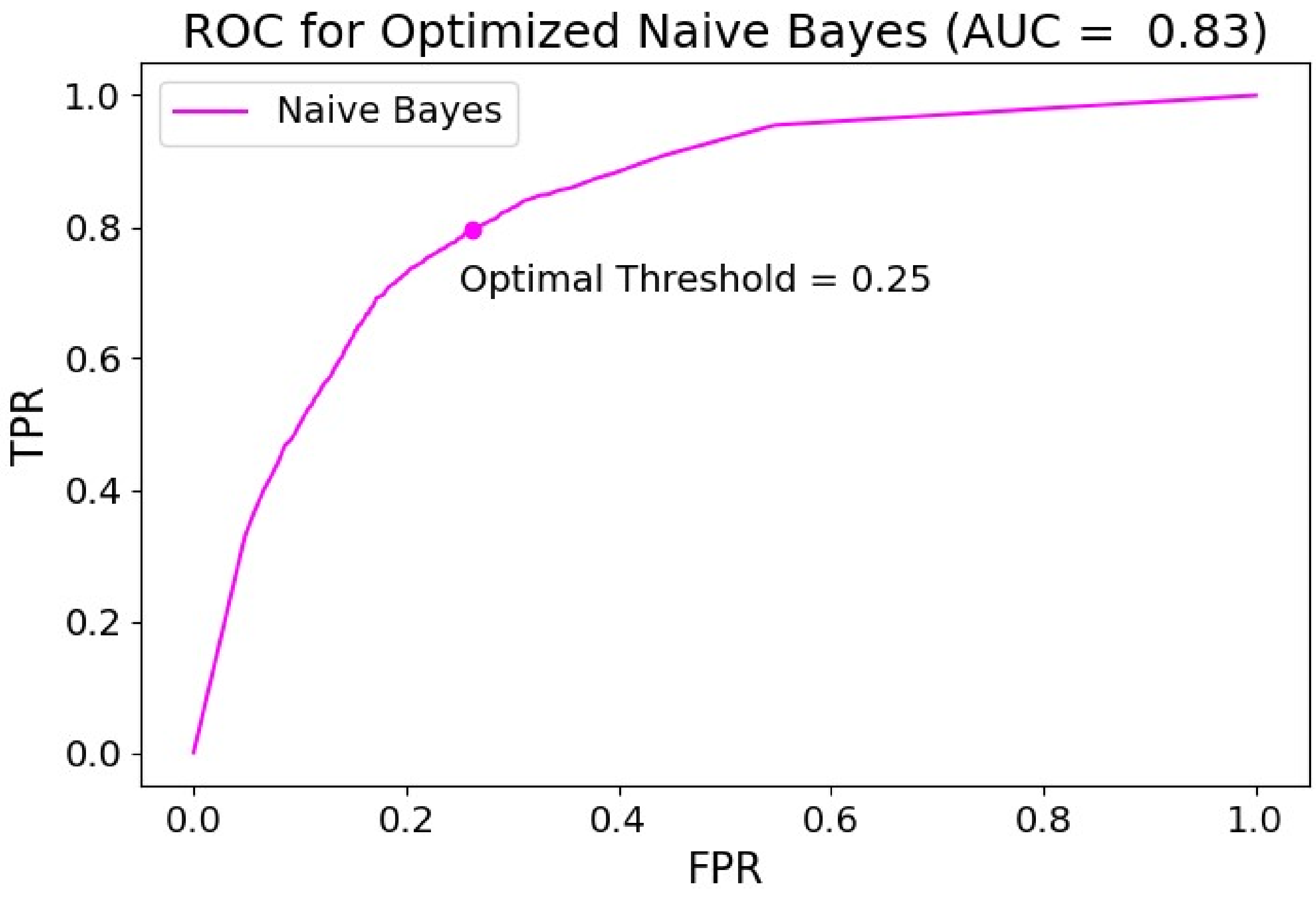
- Tested learning rates of 0.01, 0.1, 1.0, and 10
- Selected probability threshold based on largest TPR - FPR



Naïve Bayes

Metric	Value
AUC	0.83
Precision	0.28
Recall	0.80
F1	0.41
Accuracy	0.74

- No cross-validation performed
- Selected probability threshold based on largest TPR - FPR



Random Forest

- Weight-averaged entropy (ID3) used for splits
- Hyperparameters selected based on grid search
- Maximum depth of selected did not seem to influence much (though needed due to potentially high # of recursions)
- "Best" combination tended to towards max of each hyperparameter
    - *N* samples, 300 trees, 2sqrt(*D*) features
- Selected less resource-intensive set with minimal AUC tradeoff (-0.01)
    - *N/2* samples, 300 trees, 2sqrt(*D*) features)

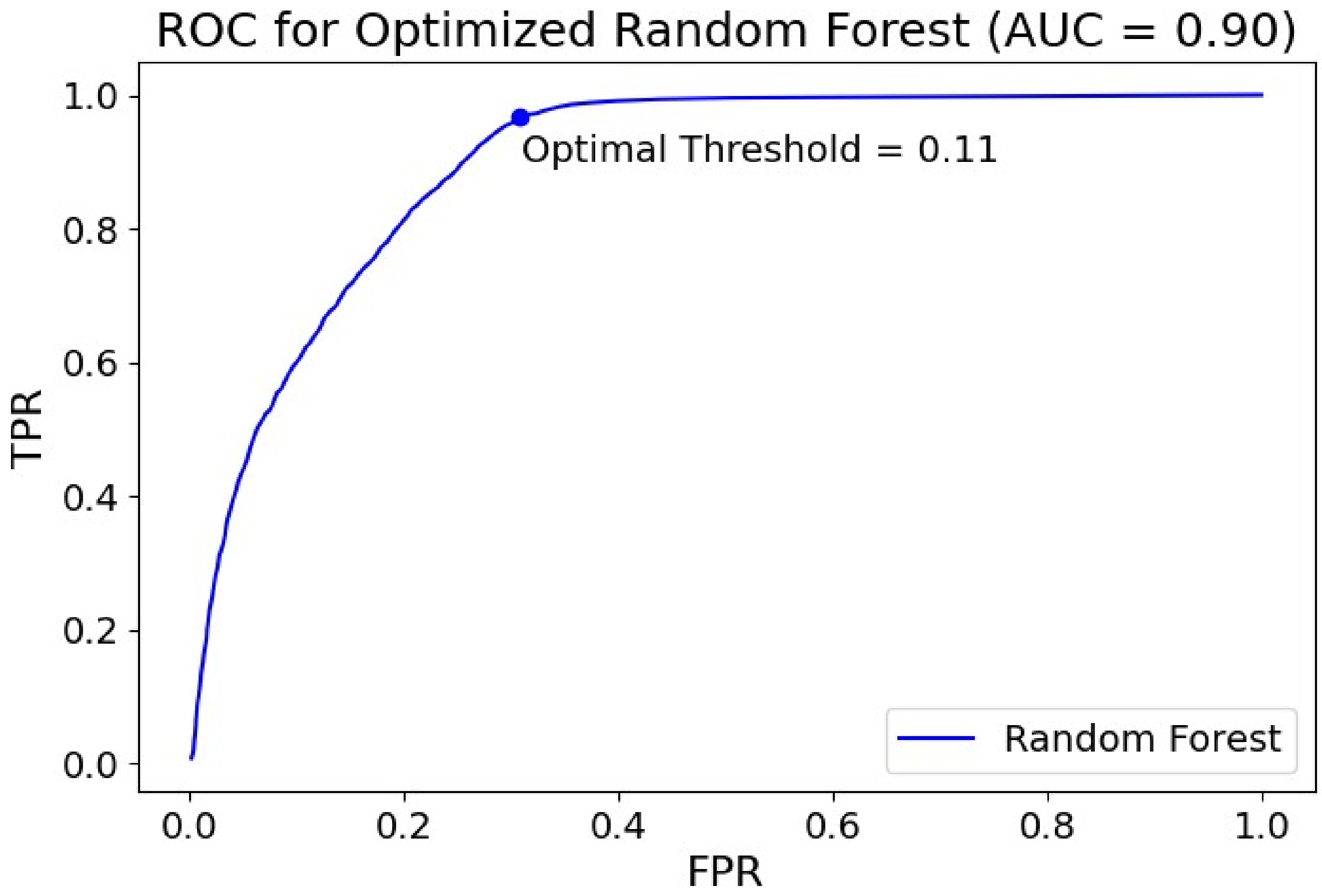
Hyperparameter	Values
Sample Size for Tree	$\sqrt{N}$, $N/2$, N
Number of Trees	50, 100, 300
Maximum Depth	100, 500, 1000, 3000
Number of Features for Split	$\sqrt{D}$, $\sqrt{D}/2$, $2\sqrt{D}$

N = number of observations, D = number of features

Random Forest

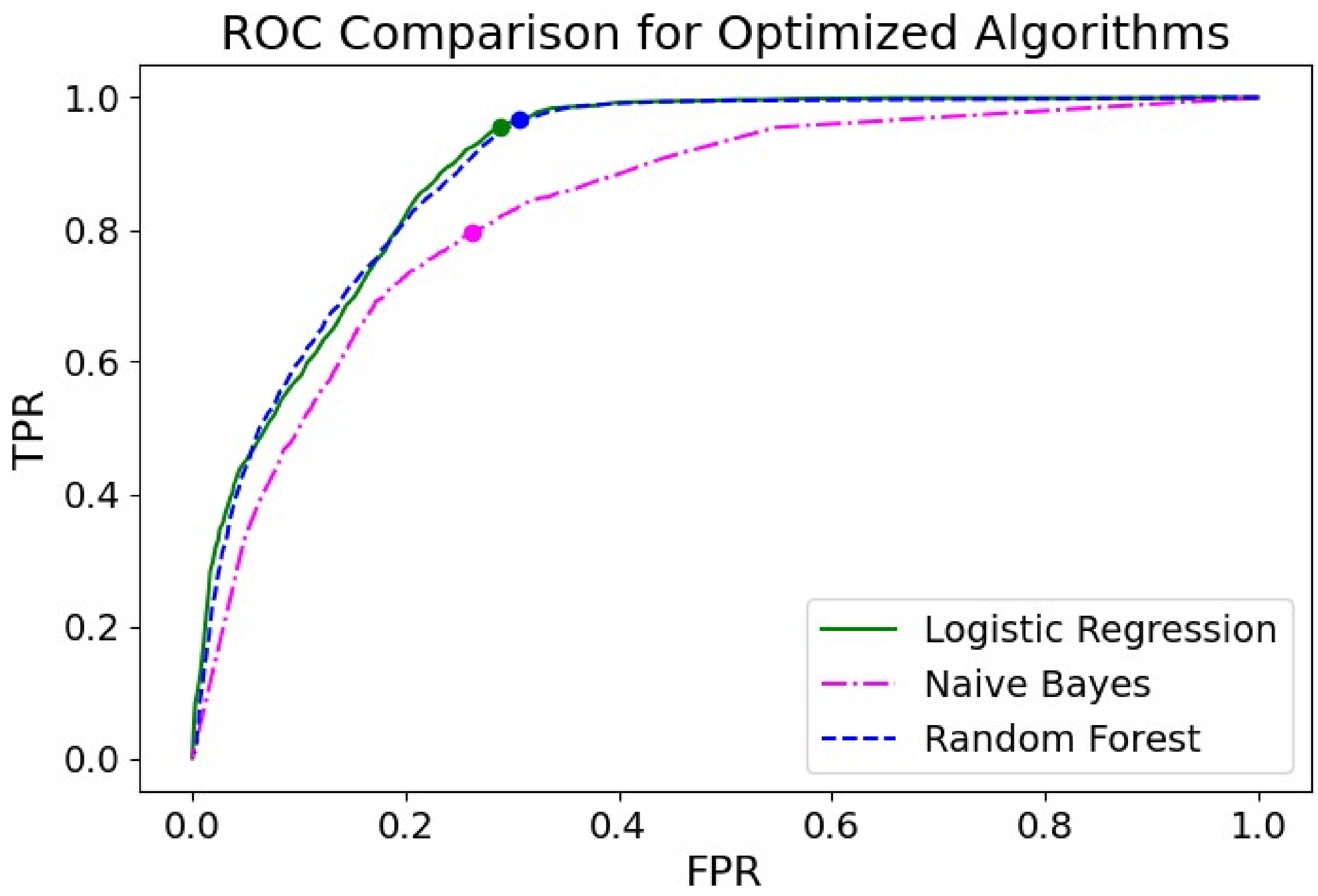
Metric	Value
AUC	0.90
Precision	0.29
Recall	0.97
F1	0.44
Accuracy	0.72

- Selected probability threshold based on largest TPR - FPR



Algorithm Comparison

	Algorithm		
Metric	LR	NB	RF
AUC	0.90	0.83	0.90
Precision	0.30	0.28	0.29
Recall	0.96	0.80	0.97
F1	0.45	0.41	0.44
Accuracy	0.74	0.74	0.72



Conclusions

- Achieved comparable AUCs to Moro et al. (~0.8 - 0.9)
- Logistic Regression would likely be best implemented in practice based on algorithms tested
  - Less resource intensive than Random Forest
  - Interpretable
- Probability thresholds and other classification metric tradeoffs could be optimized based on stakeholder preferences (e.g. may favor more TPs or fewer FPs)

Future Directions

- Rolling validation (as demonstrated by Moro et al.) to simulate model performance/tweaks over time
- Apply models to datasets from different sources, domains, timeframes (e.g. banks from different countries, more recent data)
- Different techniques to address class imbalance (e.g. SMOTE)
- Different algorithms:
  - Ensembles (gradient-boosted trees)
  - More sophisticated neural network architectures

Bibliography

1. Moro, Sérgio et al. “A data-driven approach to predict the success of bank telemarketing”. *Decision Support Systems*, vol. 62, 2014, pp. 22-31, https://doi.org/10.1016/j.dss.2014.03.001.
2. Zhu, Zining et al. “Semi-supervised classification by reaching consensus among modalities”. ArXiv. 2018. https://doi.org/10.48550/arXiv.1805.09366
3. Yoon, Jinsung et al. “ToPs: Ensemble Learning with Trees of Predictors”. *IEEE Transactions on Signal Processing*, vol. 66, 2017, pp. 2141-2152, https://doi.org/10.48550/arXiv.1706.01396
4. Pagano, Tiago P. et al. ”Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods”. *Big Data Cogn. Comput*., vol. 7, 2023, https://doi.org/10.3390/bdcc7010015
5. Chawla, N.V. et al. ”SMOTE: Synthetic Minority Over-sampling Technique”. Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321-357, https://doi.org/10.48550/arXiv.1106.1813


THANK YOU