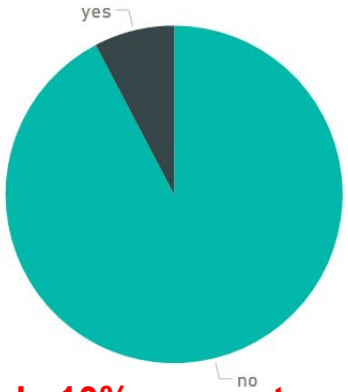
Analytics in Banking

Group 10:
Antoni Skoraczynski | Khoa Ngo | Oleksandra Syniagovska |
Muhammad Iqbal Bin Heru Wirasto | Jason Michael Sander Egan

Description of the data

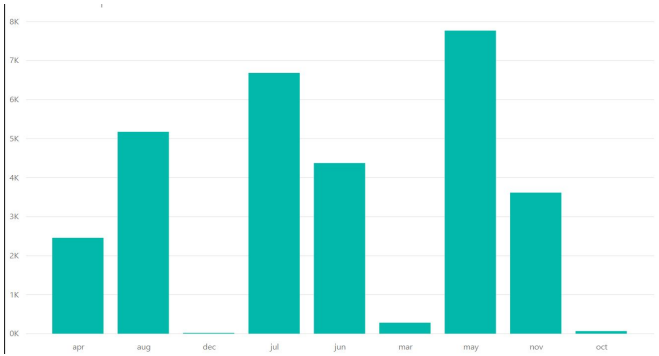
- **30436** responses, 14 questions:
  - Individual clients information (e.g. age, job, maritial status etc.
  - Interaction with marketing campaign (e.g. method of contact, frequency of contact etc.)



Only 10% are customers that opened a deposit.

27%

of responses are incomplete.



Skewed distribution within classifiers, e.g. campaign month

Linear regression models

Probability as a continuous variable

Regressor	Kaggle score (Full test data)	Conclusion
1. Age, job, marital status, previous campaign outcome, previous contact, loan/house loan history, frequency of contact during the current campaign	0.57033	Both the **loan/house loan history** and **previous contact** variables were excluded as insignificant
2. Age, job, marital status, previous campaign outcome, loan history, frequency of contact during the current campaign	**0.57029**	**Loan history** was excluded as insignificant
3. Age, job, marital status, previous campaign outcome,frequency of contact during the current campaign	0.57351	

Logistic regression models

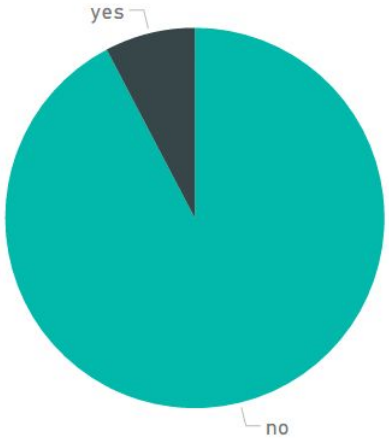
Classification problem

Classifiers	Kaggle score (Full test data)	AIC	Conclusion
1. Job, marital status, previous campaign outcome, previous contact, credit default status,frequency of contact during the current campaign.	**0.55282**	16299	
2. Job, education, previous campaign outcome, previous contact, frequency of contact during the current campaign, default history	0.55494	15572	Different job parameter and previous outcome parameter
3. Job, marital status, previous campaign outcome,frequency of contact during the current campaign, default history	0.56301	16206	
4. Job, marital status, previous campaign outcome,frequency of contact during the current campaign	0.56926	15544	

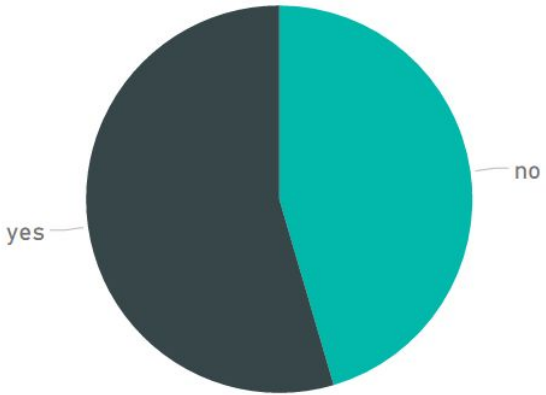
Tree model

Classification problem

- **S**ynthetic **M**inority **O**ver-sampling **TE**chnique



Initial sample:
7.7% of joined customers

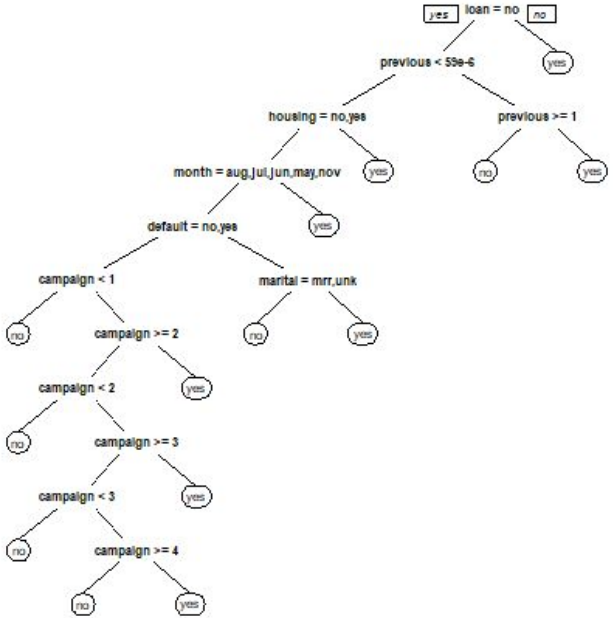


SMOTEd sample:
54.5% of ”joined” customers

Tree model

Classification problem

Loan history, **previous contact** with the bank and **frequency a person is contacted** during the current campaign is crucial for a deposit choice.



Advantages:

- Improved Sensitivity

	Reference	
Predicted	**Yes**	**No**
Yes	**117**	708
No	351	4911

TP > 0

Disadvantages:

- **58** nodes - **too complex** to be interpreted
- Still skewed classifiers
- Relatively **low predictive score**.

2.43 compared to an average 0.55 score

Conclusion

- **Logistic regression** performed the best
  - Interpretable (albeit not as readily as a linear model)
  - High predictive power
- **The best model** achieved a log loss of 0.55282/0.54204 (Private/Public).
- **Ordinary Least Squares** had the second best predictive power.
- **Tree models** were not as accurate as the other two models.
- **Previous outcome** was the most significant predictor of the marketing effectiveness
- **Loan history** and **frequency on contacting during the current campaign** also played a major role