Marketing to Banking Customers

Background

- A European bank is looking to increase its revenue
- It would like to do this by having customers buy a long term deposit product

Background

- The bank wants to know which current customers to market this product to
- We have the customer list and results from a earlier telephone marketing campaign

Caveat

- Prior marketing campaign was only directed to 14% of the customer list

Data

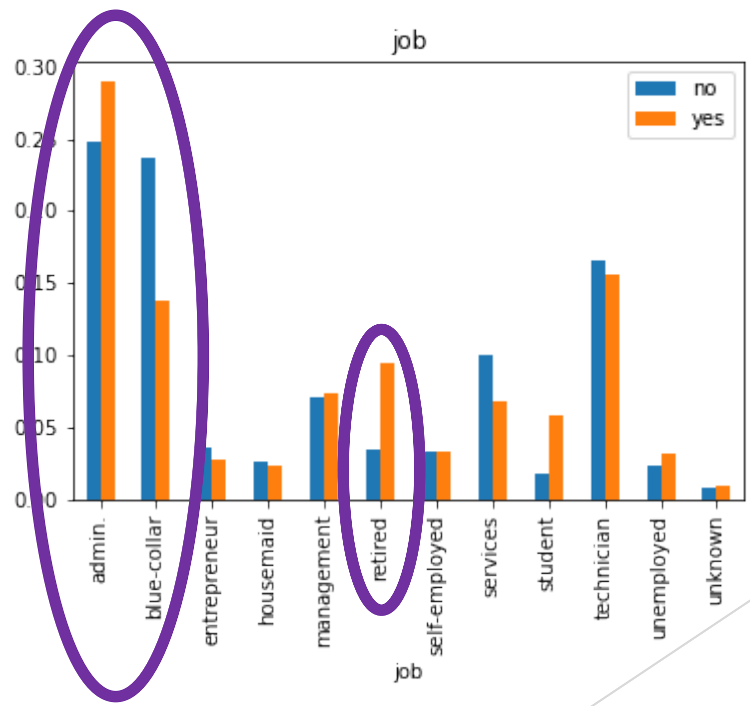
Data

- 32,154 customers
- 15 features
  - 8 concern prior marketing campaign
- Features about customers:
  - Age, job, marital status, education, whether in default, whether has a personal loan, whether has a housing loan

Features

- None of the features predict the results on their own
- Some feature values *look* predictive

Proportions of "yes" in a given job type



Design

Design

- Try to predict who will buy, given what we know about who bought before
- It’s okay to have people who didn’t buy in the previous campaign
  - 86% were not contacted in previous campaign

Model

Final Model

- Random Forest on customer features
  - That is, does not include features about the earlier campaign
- Change threshold

Final Model

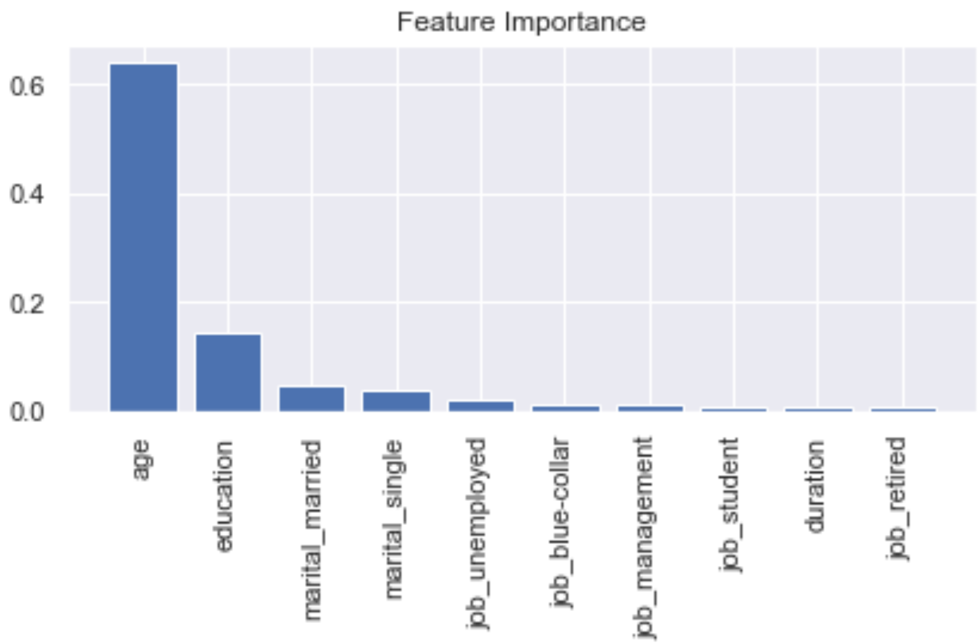
	Predicted No Sale	Predicted Yes Sale
Actual No Sale*	26	5680
Actual Yes Sale	4	721

* "Actual No Sale" includes people not marketed to

High False Positives

- Pro:
  - Only 14% of customer base was previously marketed to
  - The model may capture people who will respond positively to a new marketing campaign
- Con
  - Some of the false positives will be people who were previously marketed to and didn't buy

Top Ten Most Important Features



Future Work

Future Work

- Look at the important features to find the values that are most predictive
  - E.g.
    - Particular age range
    - Particular education level
- Use model to look at prior marketing campaign
  - Is there feature/value combination we should not market to?

Thank You!

Credits

- https://www.kaggle.com/rashmiranu/banking-dataset-classification?select=new_train.csv

Photo by Etienne Martin on Unsplash
Photo by Christian Wiediger on Unsplash
Photo by Robert Bye on Unsplash
Photo by Jonathan Cooper on Unsplash
Photo by Ibrahim Boran on Unsplash

Appendix

Age

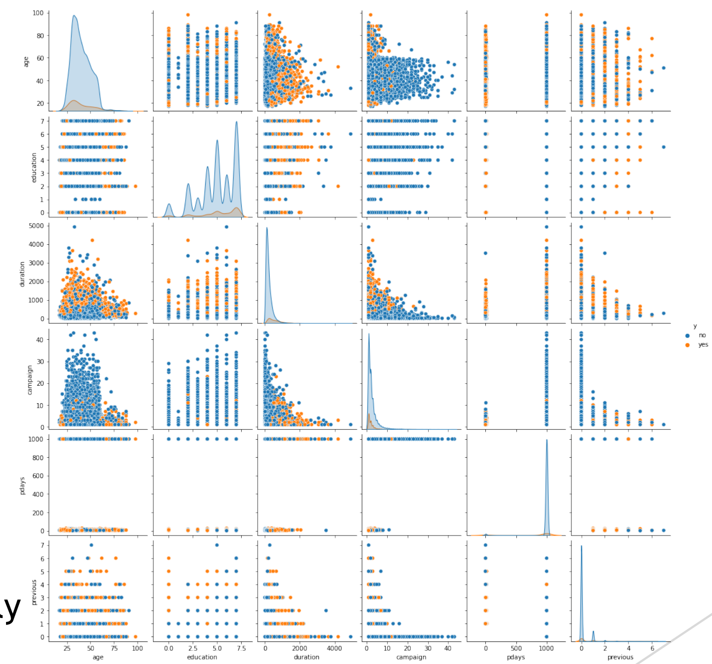
Education

Duration in seconds of last contact during last campaign

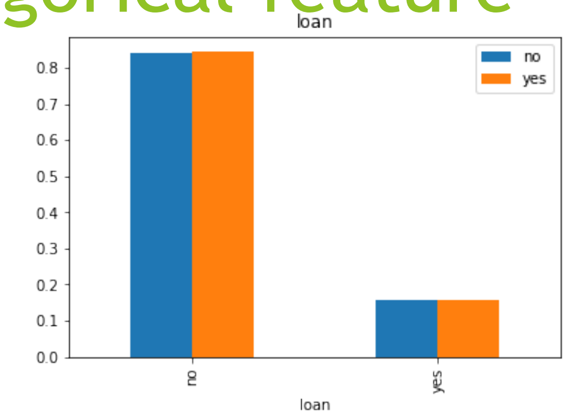
Number of times contacted during last campaign

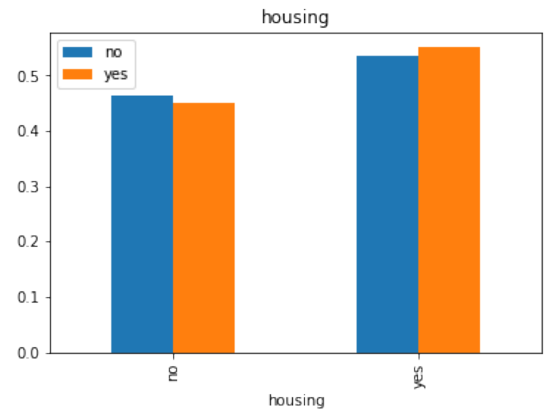
Number of days passed since contacted for prior campaign; 999 days means no contact

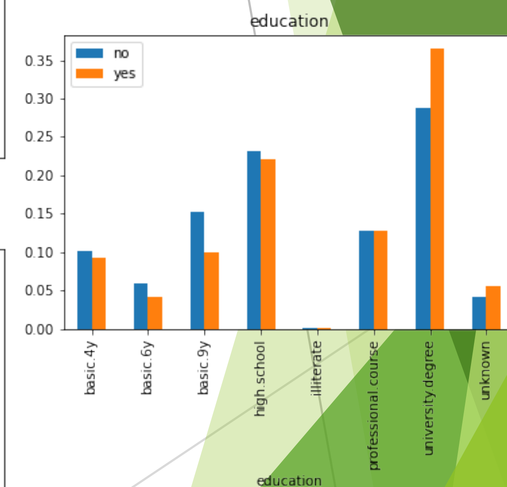
Number of times this client was contacted previously

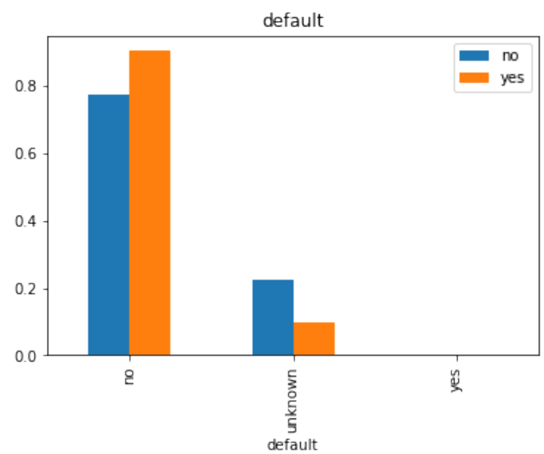


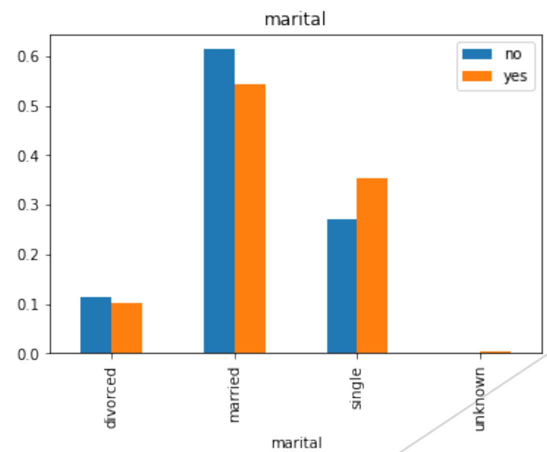
Comparing proportions of “yes” for each categorical feature



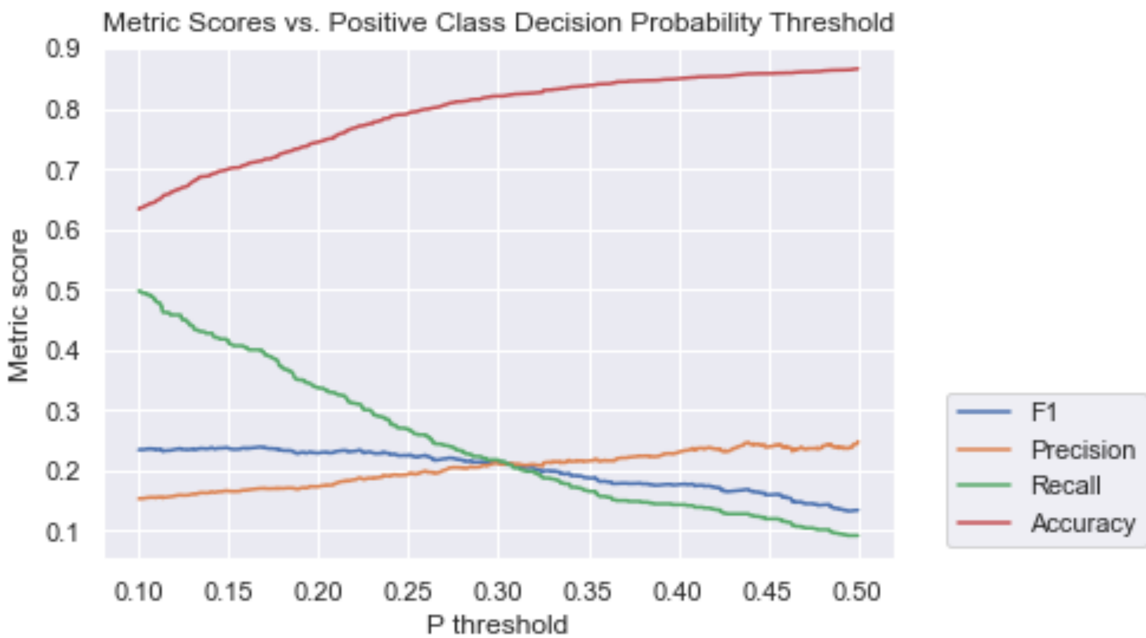








Without marketing variables



F1-max at threshold = 0.172

Random Forest without Marketing Variables – Training Data

	Predicted No Sale	Predicted Yes Sale
Actual No Sale	5493	213
Actual Yes Sale	659	66

The false negatives (Actual yes/predicted no) are higher than would prefer, and the true positives are poor, hence the modification

Recall: 0.091

Random Forest without Marketing Variables and Threshold – Training Data

	Predicted No Sale	Predicted Yes Sale
Actual No Sale	19	5687
Actual Yes Sale	13	712

The false negatives (Actual yes/predicted no) are much better, and the true positives are good. The false positives include a large number of people who weren't contacted in prior campaigns

Recall: 0.982

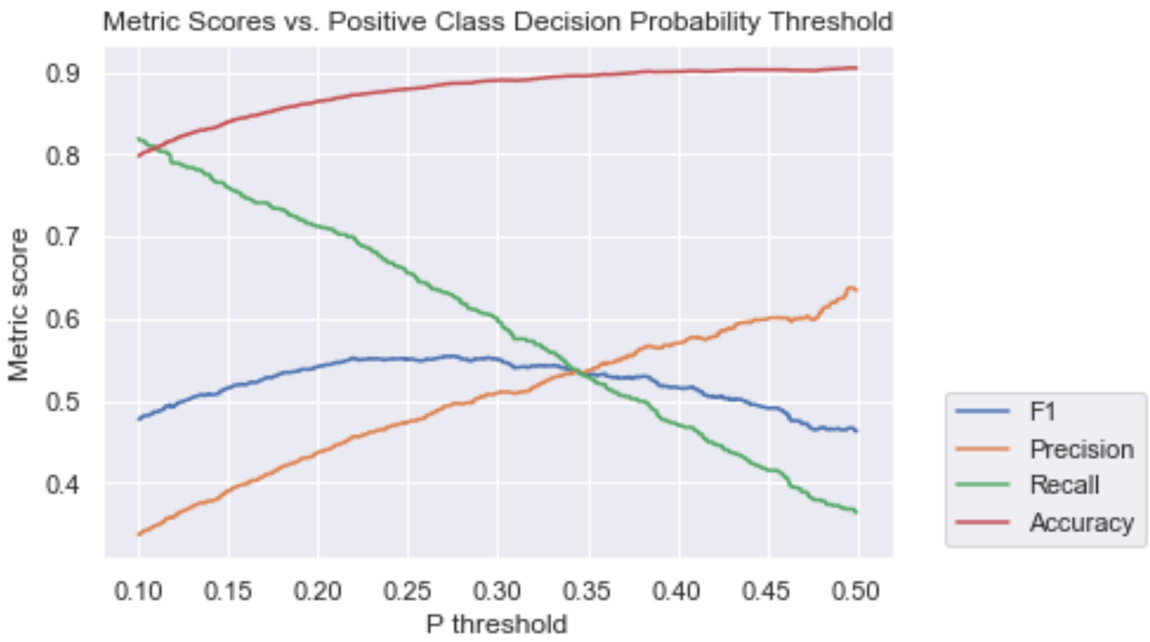
Random Forest with Marketing Variables

	Predicted No Sale	Predicted Yes Sale
Actual No Sale	5553	153
Actual Yes Sale	464	261

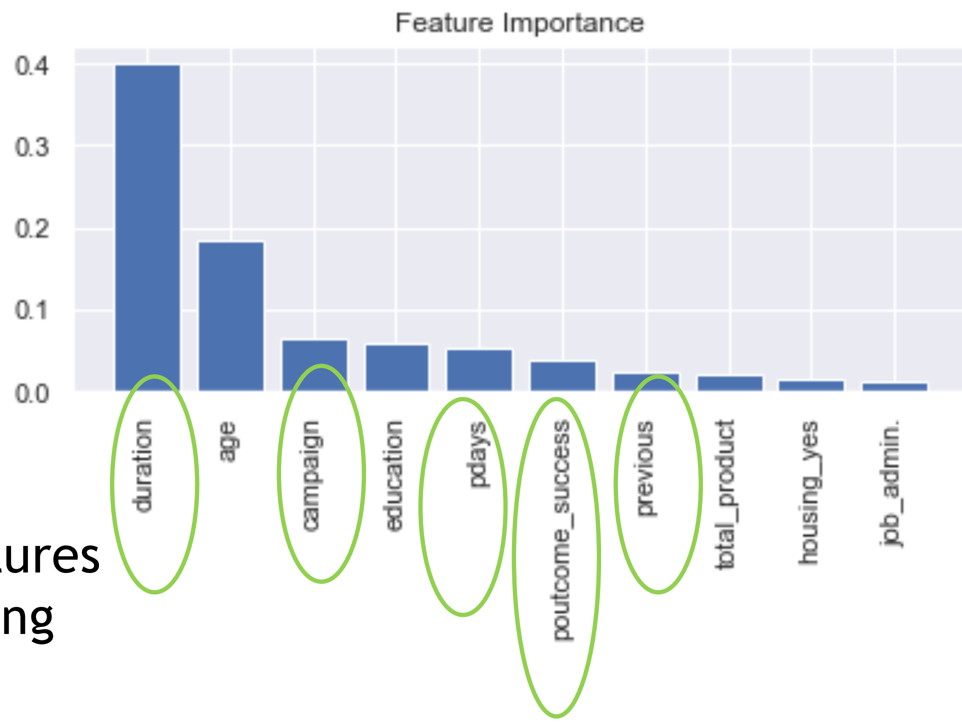
The false negatives (Actual yes/predicted no) are higher than would prefer, hence the modification

Recall: 0.360

With marketing variables



Feature Importance: with Marketing Variables



Circled features are marketing related

Logistic Regression on Oversampled Data

	Predicted No Sale	Predicted Yes Sale
Actual No Sale	5142	564
Actual Yes Sale	240	485

This has good recall and fewer false negatives (Actual Yes/Predicted No) than other models, but the modified random forest did still better

Recall: 0.669

Simple Decision Tree, Max Depth = 4

	Predicted No Sale	Predicted Yes Sale
Actual No Sale	5570	136
Actual Yes Sale	447	278

The false negatives (Actual yes/predicted no) are too high

Recall: 0.383