What page? and When?: Customers and Non-customers

Daniel Thoren s144222

Introduction

Customers and Non-customers browse Danske Bank's website differently.

How can browsing behaviour be used to predict whether a visitor is a customer or not?

Features

By examining…

- Which page visitor is on
- Pages seen during visit

…we can predict if this visitor is a customer of the bank.

Use Cases

- Improved **customer acquisition** by showing relevant content to non-customers.
- Improved **customer retention** by showing relevant content to existing customers.

Feature Extraction

The above mentioned features are derived from:

page_*.csv

- iscustomer
- pagelocation
- pagesequenceinsession

One-Hot-Encoding
URLs are extracted from **pagelocation** and converted to dummy variables.

Performance

A range of classification models were trained and tested on the dataset. After reviewing the results, **Random Forrest** came out on top.

	Pos	Neg	
Pos	**277** TP	**74** FP	**78.9 %** PPV
Neg	**18** FN	**210** TN	**92.1 %** NPV
82.9 % Accuracy	**93.9 %** Sensitivity	**73.9 %** Specificity	

Feature Importances

Although the majority of the prediction is based the current page, knowing how many pages deep the visitor is, improves accuracy significantly. The importance of location varies from page to page:

Which page
Number of pages

39.7% of the model's decision making progress comes from which page you are on.

Some pages, such as those only accessible to customers - or only interesting to potential customers - give a clear indicator.

The URLs on the left are microscopic portion of pages that most-clearly-indicate the visitors customer status.

Visiting one of these URLs **partially** predicts whether the visitor is customer or not.

Whether it's the **first page** you visited or the **twenty-first** plays a role too.

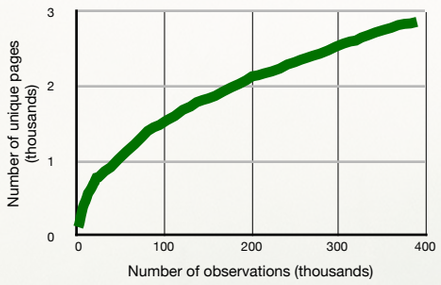
5.2 %	.dk/privat/find-hjaelp
4.6 %	.dk/privat/logoff-netbanken
3.5 %	.dk/privat
2.2 %	.fi/sinulle/logoff-verkkopankki
1.1 %	.no/privat/logoff-nettbanken
1.1 %	.se/privat
1.1 %	.no/privat/faa-hjelp
1.0 %	.fi/sinulle/asiakaspalvelu
0.9 %	.se/privat/kundservice
0.9 %	.co.uk/personal
0.9 %	.fi/sinulle
0.8 %	.se/foretag/kundservice

Findings

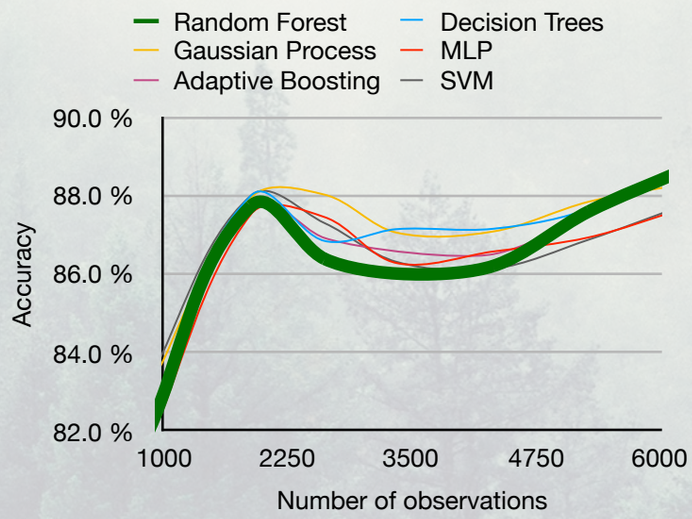
Training and Pre-processing

Pre-processing data to get fewer unique sites than observations is key to training a classifier.

As Danske Bank's website has many many subpages, one-hot-encoding results in very sparse matrices.



Scalability



Future Work

- Recover unescaped data instead of discarding
- Group URLs by pre-fix substring
- Train on full dataset

Photo by Jaymantri