# C&C pilot project

Anton Örn Karlsson

Statistics Iceland

April 16th 2020

# Overview

- A brief background
- Our project
- Technical aspects
- Future plans
- Lessons learned

Statistics Iceland

# Background

- Statistics Iceland
  - Founded in 1913
  - Member of EFTA (European Free Trade Association)
  - Member of the European Statistical System
    - Produces most of the same official statistics as other European nations
  - Approximately 110 full time employees (excl. interviewers)
  - Situated in Reykjavik, the capital of Iceland
- Policy for 2020-2025
  - Among the goals:
    - **Coordinate methods and automatize processes**
    - Implement solid technical infrastructure
    - Support innovation and the acquisition of knowledge

# The Icelandic C&C project

- Use of machine learning techniques in social statistical production
- Automatic coding
  - Occupation
  - Industry
- Open responses gathered by telephone interviewers
  - EU-SILC
  - LFS
  - etc.

Statistics Iceland

# The goals

- Reduce the needed manpower for coding
- Increase the speed of coding
  - Increased timeliness of results
- A case for increased use of ML in official social statistics
  - Re-usable code for use within Statistics Iceland (and beyond)

# The basic plan

- Available coded data
  - LFS from 2003
  - EU-SILC from 2004
- Train models on past coded data
  - More than one model in order to make comparisons
  - Select relevant indicators for models
- Implementation

Statistics Iceland

# Technical aspects(1)

- The project will be completed using R
  - gitlab for version control
  - All data in MSSQL tables
- Main libraries
  - tidyverse
    - dplyr, dbplyr, ggplot2, modelr, purrr, stringr, etc.
  - tidytext
  - caret
  - RODBC / DBI / odbc
  - tm
  - gmodels
  - tokenizers.bpe
- Data cleaning has taken a long time
  - learning curve for myself
  - Challenging to use text classification techniques for Icelandic

Statistics Iceland

# Technical aspects(2)

```
ord <- c("og","í","á","um","við","með","að","til"
         ,"sem","sjá","því","vera","frá","gera"
         ,"hjá","ofl","osfrv","upp","út")

#storf_corpus_clean <- storf_corpus %>%
  #tm_map(content_transformer(tolower)) %>%
  #tm_map(removeNumbers) %>%
  #tm_map(removeWords, ord) %>%
  #tm_map(removePunctuation) %>%
  #tm_map(stemDocument) %>%
  #tm_map(stripWhitespace)

#storf_dtm <- DocumentTermMatrix(storf_corpus_clean)
```

# Technical aspects(3)

- Models being considered
  - Naive Bayes
  - Support vector machines
  - Random forest
  - XGBoost (Extreme Gradient Boosting)
  - Deep learning (if possible)
- Other ideas would be greatly appreciated
  - FastText

Statistics Iceland

# Technical aspects(4)

- How to compare models?
  - Confusion matrix
    - Accuracy
- Again, any ideas would be greatly appreciated
  - I will surely go through the papers and presentations from the sprint to get ideas

Statistics Iceland

# (Very) Preliminary results

- Three digit occupation coding
  - Only using data from the LFS in 2017-2019
  - Naive Bayes
    - Accuracy: 0.7372
  - Support Vector Machines
    - Accuracy: 0.0577
    - Everyone should be coded as a shop salesperson (522)
    - Maybe a class imbalance problem?
    - Back to the drawing board!

Statistics Iceland

# Implementation plan

- Final model, ready
- Set up a CI/CD solution on gitlab
  - This solution is still being developed
  - Hopefully it will be ready when the ML model is ready
- Regular implementation of the ML model on new un-coded data
  - Also possible for experts working with the data to run the model
- Additional resources
  - The code will be shared on internal gitlab
  - Metadata for users detailing how classification is being conducted
  - Presentations and internal material for Statistics Iceland

Statistics Iceland

# Future look

- Generic solution for other classification and coding processes
  - Education
  - Census classifications
  - New statistics based on text data
- Implementation within data collection software
  - Very far in the future
  - Something I heard from other presenters in the sprint

Statistics Iceland

# What I have learned from this virtual sprint

- Sentimental analysis of twitter for Iceland
- Be careful what you promise - are there any low hanging fruits?
- Time is of the essence
  - And know-how and interest
- The field of ML moves very quickly
  - Lots of new developments
  - That is why this cooperation is so important!

Statistics Iceland

# THE END

THANK YOU!

Statistics Iceland