Predicting Competence in a Foreign Language

Capstone Part 5: Results

Aim

Problem Statement:

- Applications for degree courses linked to European languages have fallen by almost a quarter in the past five years, and applications to other language courses have dropped by almost a fifth (source, Press Association analysis).

Q: Would it make a difference if we could predict competence in a foreign language?

Audience & Goals

Target Audience:

- Schools/boards of education who need justification for funding in targeted areas
- Students who need justification of the factors leading to success in order to continue studying languages

Goal:

- A model that can predict the achievement of competence in a foreign language on the basis of variable factors (ie. not moving to the country in question or being adopted by bilingual parents)

Approach

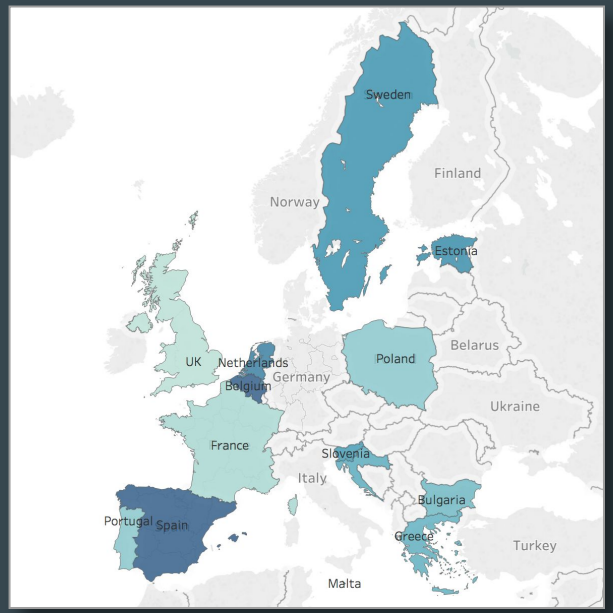
Exploratory Data Analysis

- Select data set
- Set Target
- Data analysis to understand distribution of data and correlation with the Target
- Selection of features with strongest correlation with Target
- (Removal of ones with strong co-correlations)

Modelling

- Establish baseline accuracy (0.68)
- Select classifier models (~9)
- Train models against 70% of the data
- Tune models to identify best hyperparameters for optimal model performance
- Test models against remaining 30% of the data
- Identify best performing models, score and visualise outputs (and communicate to you guys!)

Data



Source:

- European Survey on Language Competences (2012, Centre for Research on Education and Lifelong Learning)

Summary:

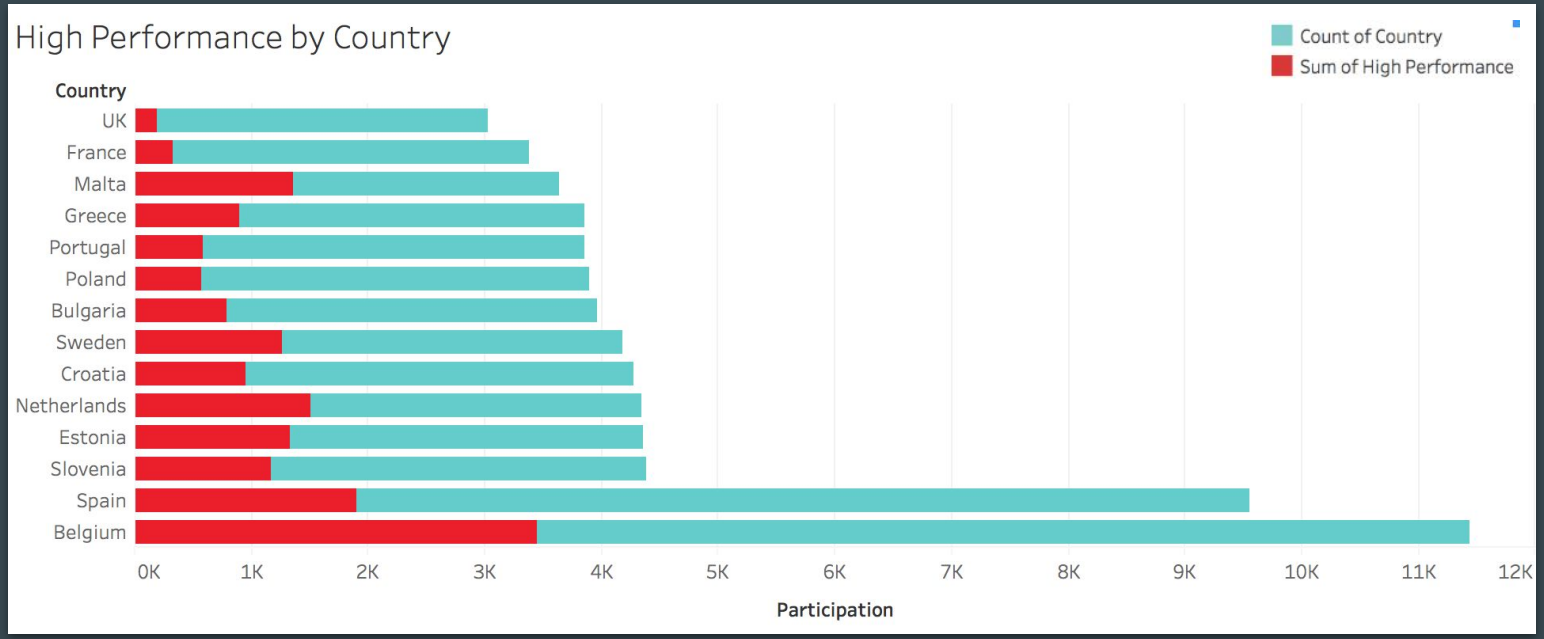
- 14 countries
- ~53k secondary school students
- 499 data fields:
  - Results of language tests in listening, reading and writing skills, scored according to the Common European Framework of Reference (CEFR)
  - Results of administered questionnaires

European Commission

Test Results (CEFR) & Classification

Level Name	Level Code	Description	Proportion of Results	Target	Proportion of Target
N/A	-A1	Below minimum level	0.16	Low (0)	0.69
Basic user	A1	Breakthrough or beginner	0.37		
	A2	Waystage or elementary	0.16		
Intermediate user	B1	Threshold or intermediate	0.17	High (1)	0.31
	B2	Vantage or upper intermediate	0.14		

Test Results (CEFR) & Classification



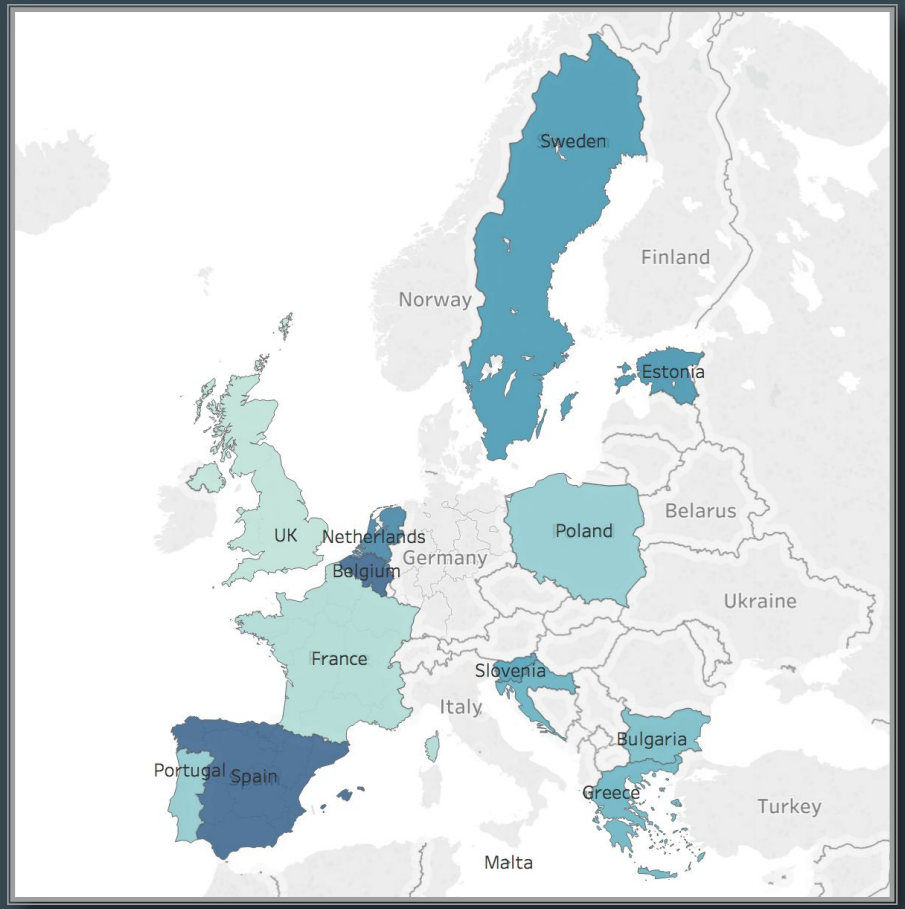
Sample Survey Questions

Language-specific:

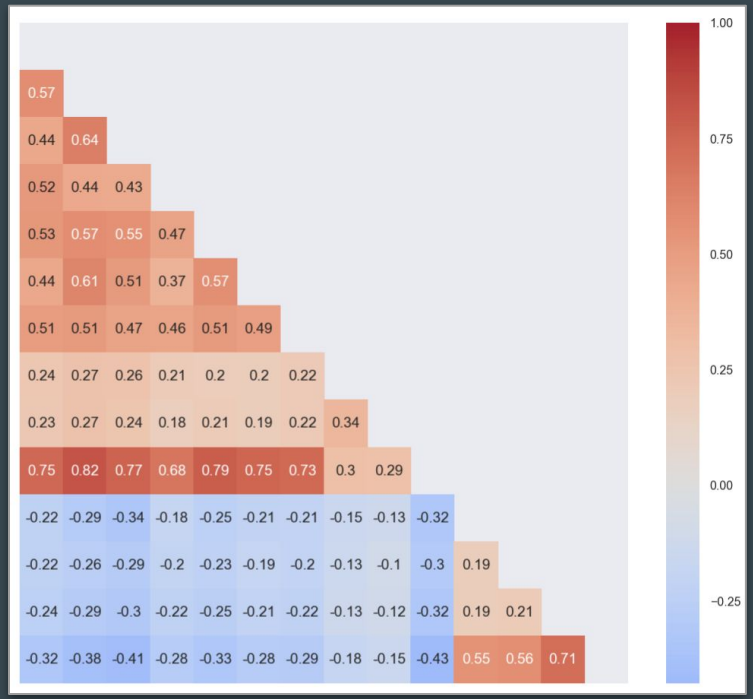
- Languages spoken at home
- Exposure to target language outside school
- Perceived difficulty/usefulness of language
- Parents' language knowledge

General:

- Time spent studying
- Class size
- Enjoyment of school subjects
- Access to IT/internet and frequency of usage
- Socio-economic factors



Correlations



Positive:

- Parents' knowledge of target language
- Perceived usefulness of language (particularly for entertainment and computing)
- Enjoyment of target language

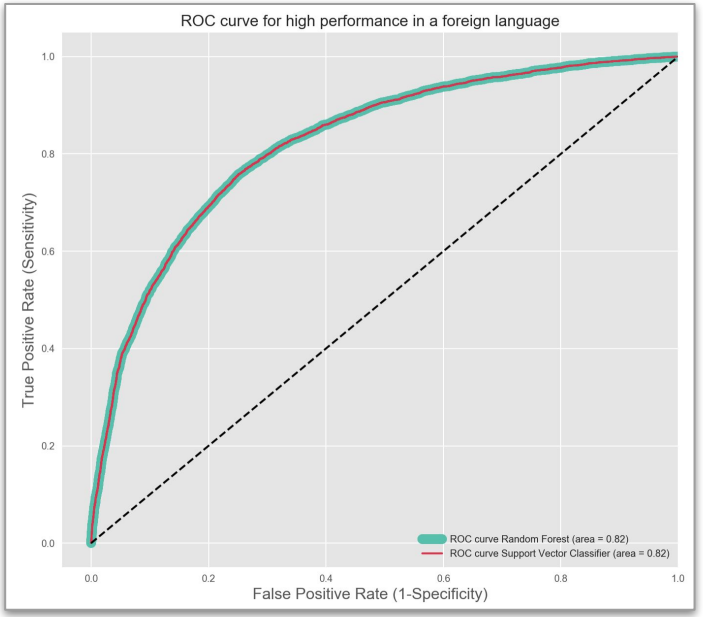
Negative:

- Perceived difficulty of learning and understanding the target language

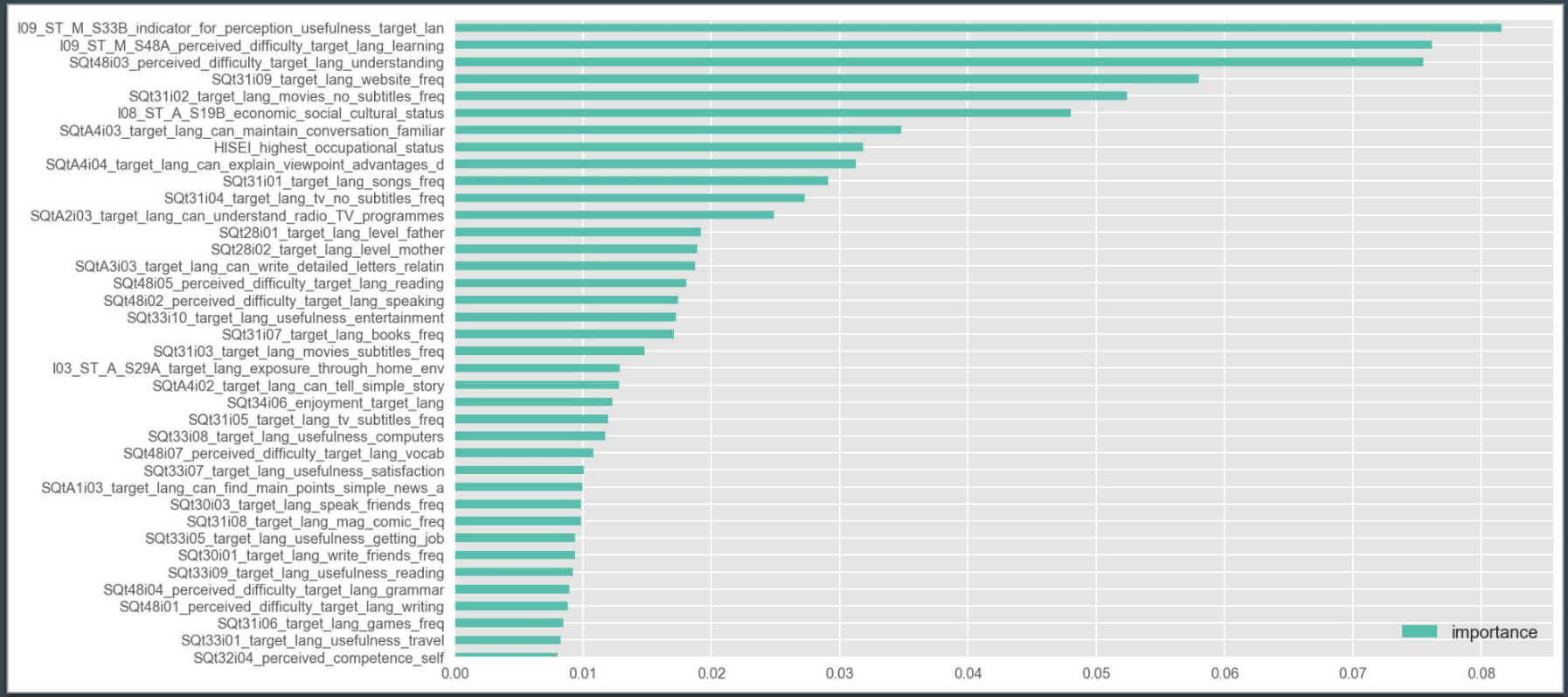
- Frequency of PC usage in classes and for study is very strongly correlated across all school subjects and non-study activities BUT not correlated with high performance

Best Performing Model

Model	Precision	Recall	Accuracy	*Baseline Accuracy*
Random Forest	0.81	0.82	0.82	*0.68*



Best Predictors



Best Predictors

Language-specific:

- Perceived usefulness (job/entertainment/IT)
- Difficulty
- Frequency of media consumption (websites/films/TV/music)
- Perceived competence (ability to understand TV & Radio/hold a conversation/write letters, etc.)
- Parents' language ability
- Enjoyment of language

General:

- Socio-economic status

Limitations

I would struggle to recommend this model for the original purpose for the following reasons:

- 49% of the dataset relates to English which could be skewing the feature importance (ie. media)
- Better at predicting low performance than high performance
- Correlation does not necessarily mean causation (does achievement make it enjoyable or vice versa?)

Next Steps

Expand modelling as follows with a view to deciding whether it can be used for a different purpose:

- high/low performance balance in dataset
- target language balance in dataset
- discipline-specific Targets (listening / reading / writing)
- multi-class modelling

Review goal to inform focus for next steps

Thank you

Liz Spiking
DSI5