Web User Demographic Characteristics Prediction

March 24th, 2023 – Group 5

Table of Contents

Using Stacking to Identify Gender and Age Based on User's Online Journey at a 82.9% Accuracy

Situation

Needs to Understand Customers

Demographic information can help businesses target their advertising more effectively. By understanding the demographics of their audience, businesses can create more targeted advertising campaigns that are more likely to resonate with their target audience.

Overview

Data Source

The dataset contains columns with unique ID for panelists, their declared gender, internet session ID, date, time, page domain, and page URL. These columns can be used to analyze online behavior and preferences, track user sessions, and identify popular content and features on specific pages.

Question

How to know the user based on their online journey?

Solutions

Approach

We created new variables such URL word frequency from the dataset and then trained the extracted data on machine learning models.

Prediction Outcome

We developed a stacking model to identify gender and age at 82.9% accuracy.

Part 1

Data Pre-Processing

Dataset Overview

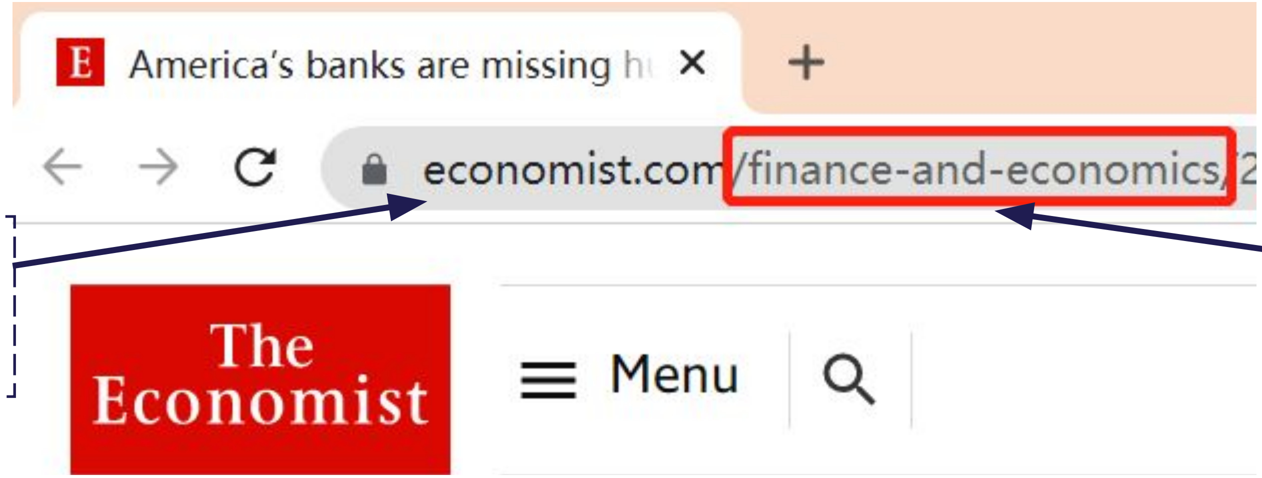
panelist_id	OS	gender_char	age_group_char	social_grade_char	Date	Time	PageDomain	PageUrl
1.83657E+18	Android 11	female	25-34	c2	2022/12/1	18:08:22	tester.userbrain.net	tester.userbrain.net/c
1.83657E+18	Android 11	female	25-34	c2	2022/12/1	13:10:03	google.com	google.com
-7.2063E+18	Windows 10	female	45-54	ab	2022/12/1	11:39:38	outlook.live.com	outlook.live.com/ma
-7.2063E+18	Windows 10	female	45-54	ab	2022/12/1	12:00:38	outlook.live.com	outlook.live.com/ma
-7.2063E+18	Windows 10	female	45-54	ab	2022/12/1	18:22:20	www.raileurope.com	raileurope.com
-7.2063E+18	Windows 10	female	45-54	ab	2022/12/1	18:58:20	hodmedods.co.uk	hodmedods.co.uk/c
-7.2063E+18	Windows 10	female	45-54	ab	2022/12/1	11:13:57	outlook.live.com	outlook.live.com/ma
-7.2063E+18	Windows 10	female	45-54	ab	2022/12/1	11:39:48	outlook.live.com	outlook.live.com/ma
-7.2063E+18	Windows 10	female	45-54	ab	2022/12/1	12:31:06	spotl.io	spotl.io/en/pu/tasks
-7.2063E+18	Windows 10	female	45-54	ab	2022/12/1	20:02:49	translate.google.co.u	translate.google.co.u

- ➔ **Raw Data: 3100** datasets in the data pack, **9** columns
- ➔ **Columns:** unique ID for panelists, gender, internet session ID, date, time, page domain, and page URL

We merge all the rows and drop empty values.
As a result, we get 26.5M rows.

Variable Formulation - Using the Information in URLs

Page Domain

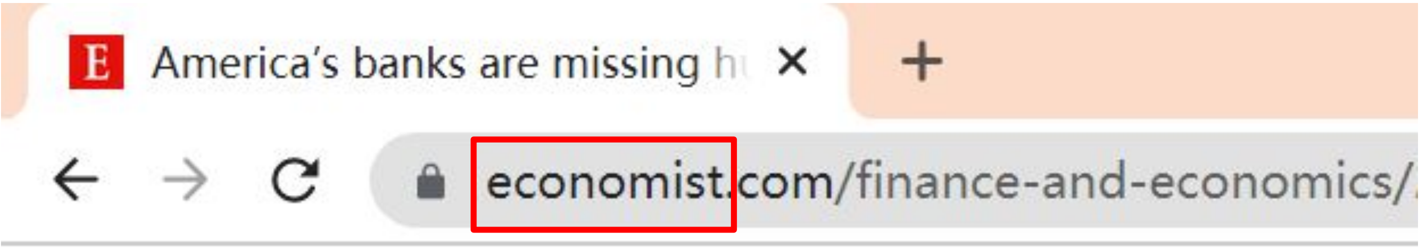


Path

The URLs contain information on webpage content

Variable Formulation - Using the Information in Page Domain: the Approach

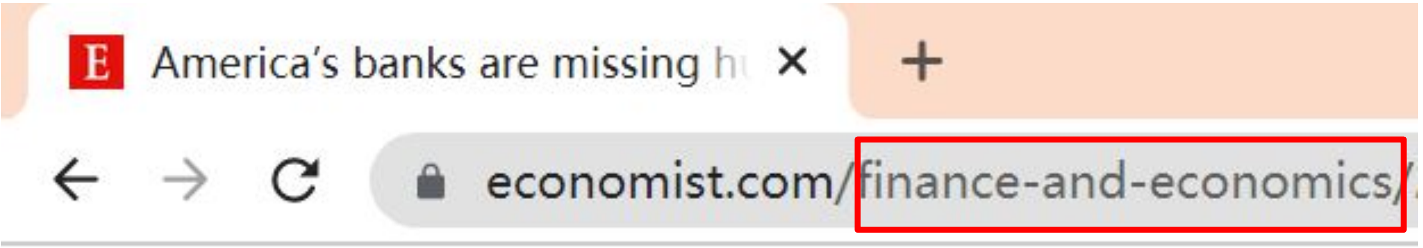
1. **Remove** "technical" parts i.e., www, .com, .uk, .org, …
2. Calculate the **domain frequency** for each user
3. Find the top **500 frequent domains** for all groups and take the **union**
4. Use **random forest** to select the top **50% important domains**



"economist"

Variable Formulation - Using the Information in Path: the Approach

1. **Replace** non-English characters with a **space**
2. **Tokenize** and **split** the words
3. Perform **stemming** to reduce words to their word stem
4. Calculate the **word frequency** for each user
5. Find the top **500 frequent words** for all groups and take the **union**
6. Use **random forest** to select the top **50% important words**



"finance"
"economic"

Variable Preparation - Producing New Data and Normalize Scaling Data

Features

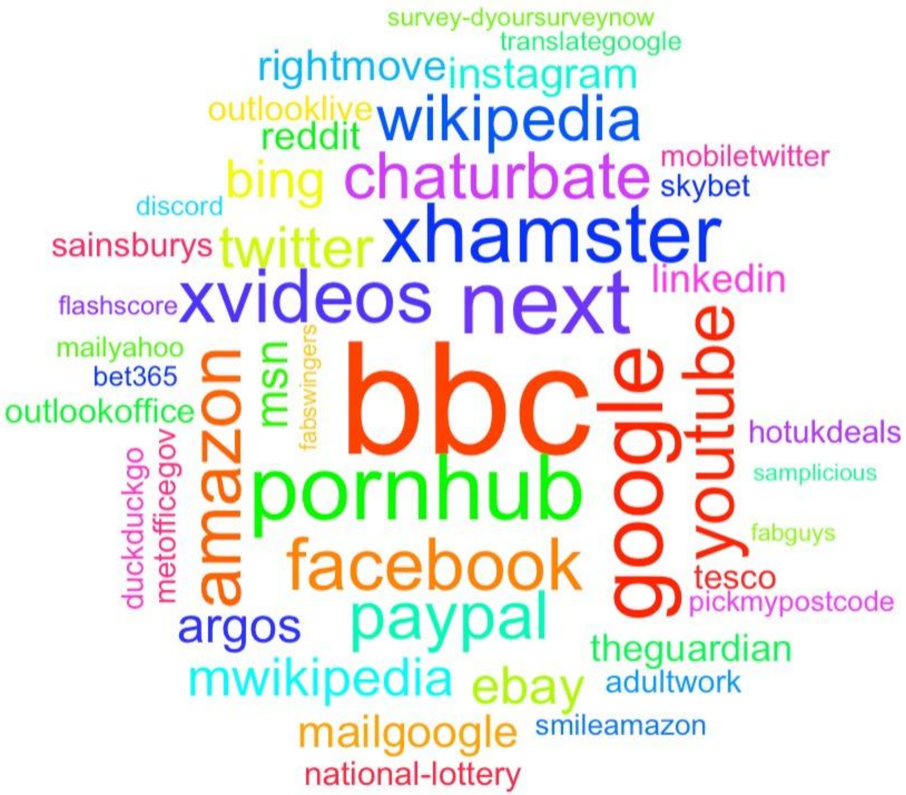
- Domain Frequency
- Words in Path Frequency
- Number of Pages Visited
- Weekend/ Weekday Ratio
- Time Window Frequency
- Visit Duration

In the dataset, features or variables have **different scales**, so we **normalise** the data using the **Z-score**

Part 2

Modelling

Inspiration – Web Importance Difference between Genders



Male

Female

As gender tends to influence website preferences, we consider web visit importance a crucial feature for our modelling.

Fitting Method - Use Cross-Fold Validation to Validate Our Models

80 / 20

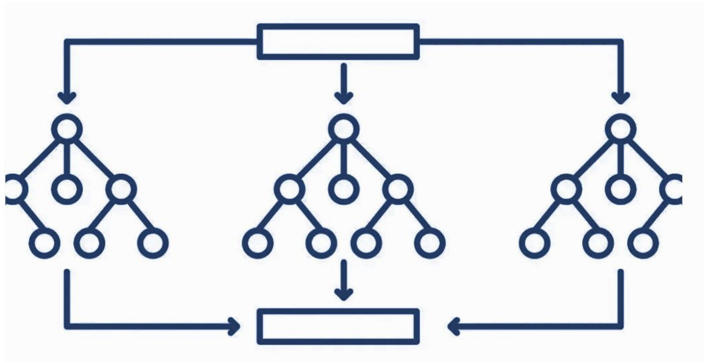
Train-test split

5-fold cross-val.

Our hyperparameter tuning method

Models Used (part I) - Random Forest, Naïve Bayesian Classifier, K-Nearest Neighbor

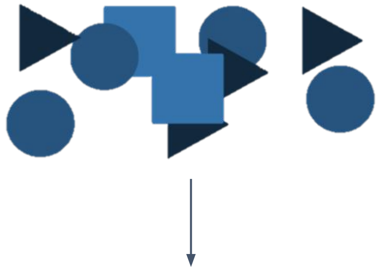
1 Random Forest (XGBoost)



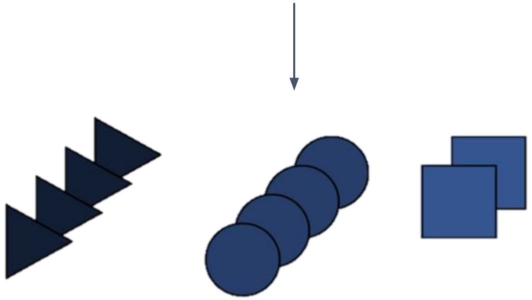
Max-depth of 5 levels

Learning rate of 0.1

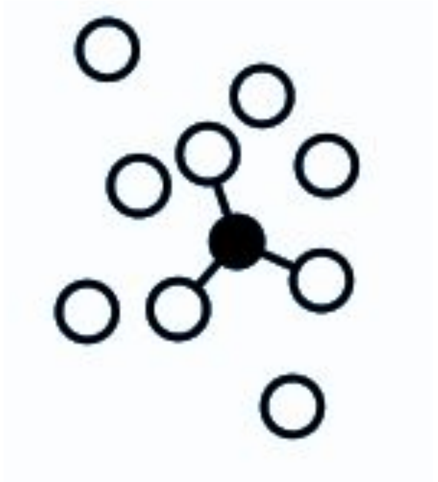
2 Naïve Bayesian Classifier



$$P(A) = \frac{P(B|A)\times P(A)}{P(B)}$$



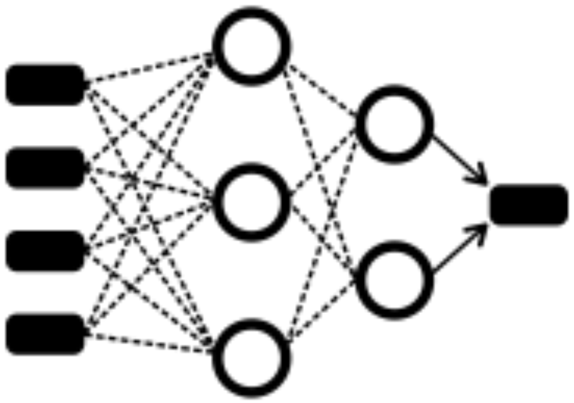
3 K-Nearest Neighbor



Optimized K = 14

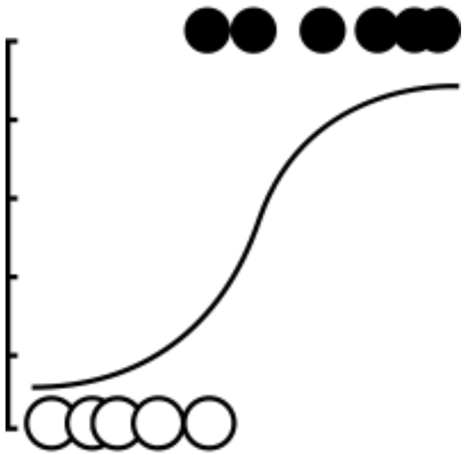
Models Used (part II) - Neural Network, GLM, Stacking

4 Neural Network

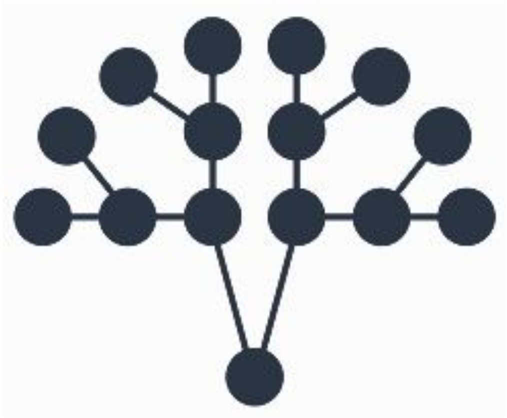


4 layers of 5 neurons

5 GLM (Logistics Regression)



6 Stacking

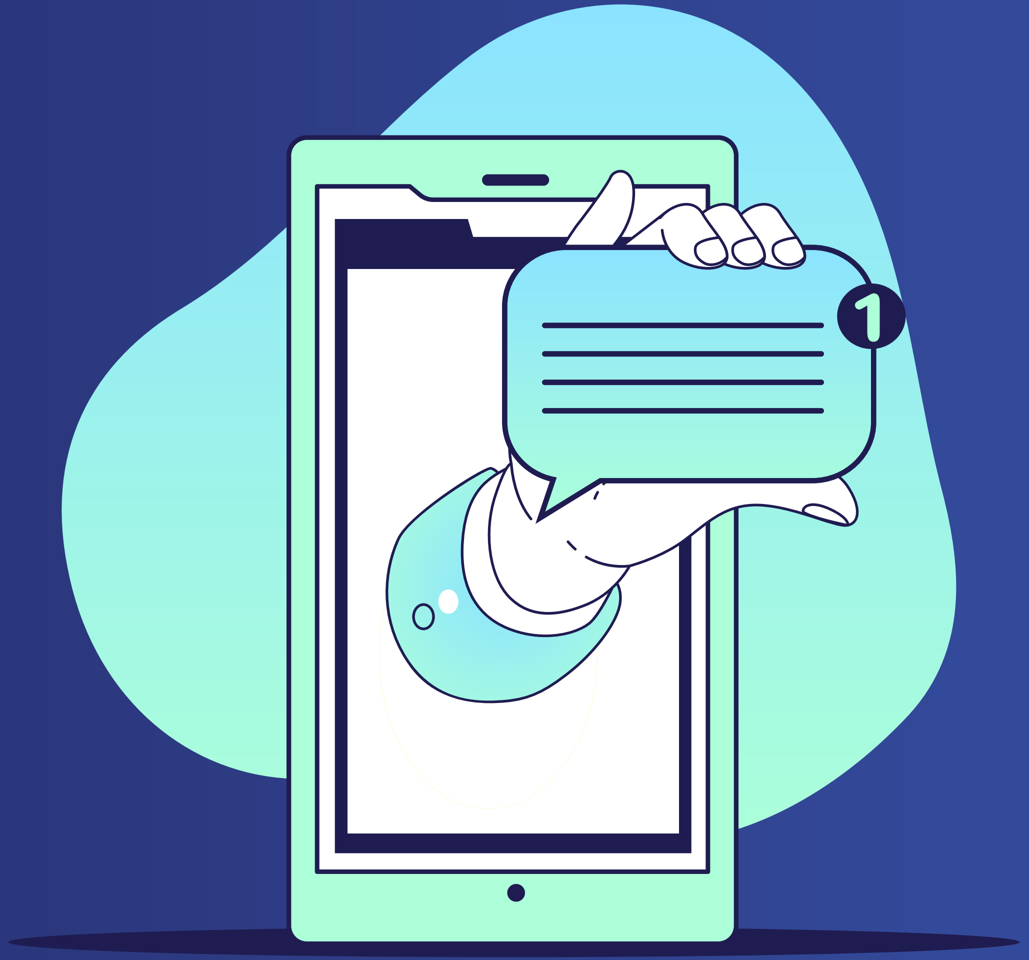


Combination of the
5 previous models

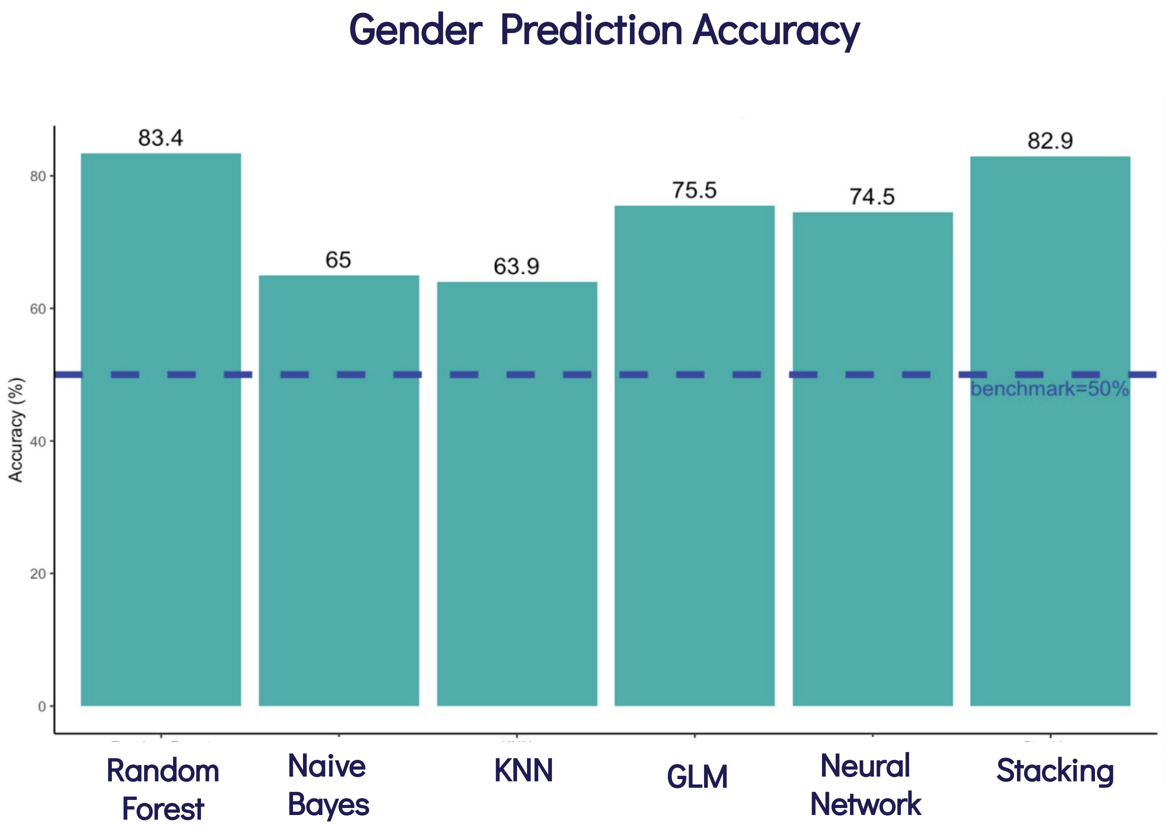
Meta-model: logistic regression

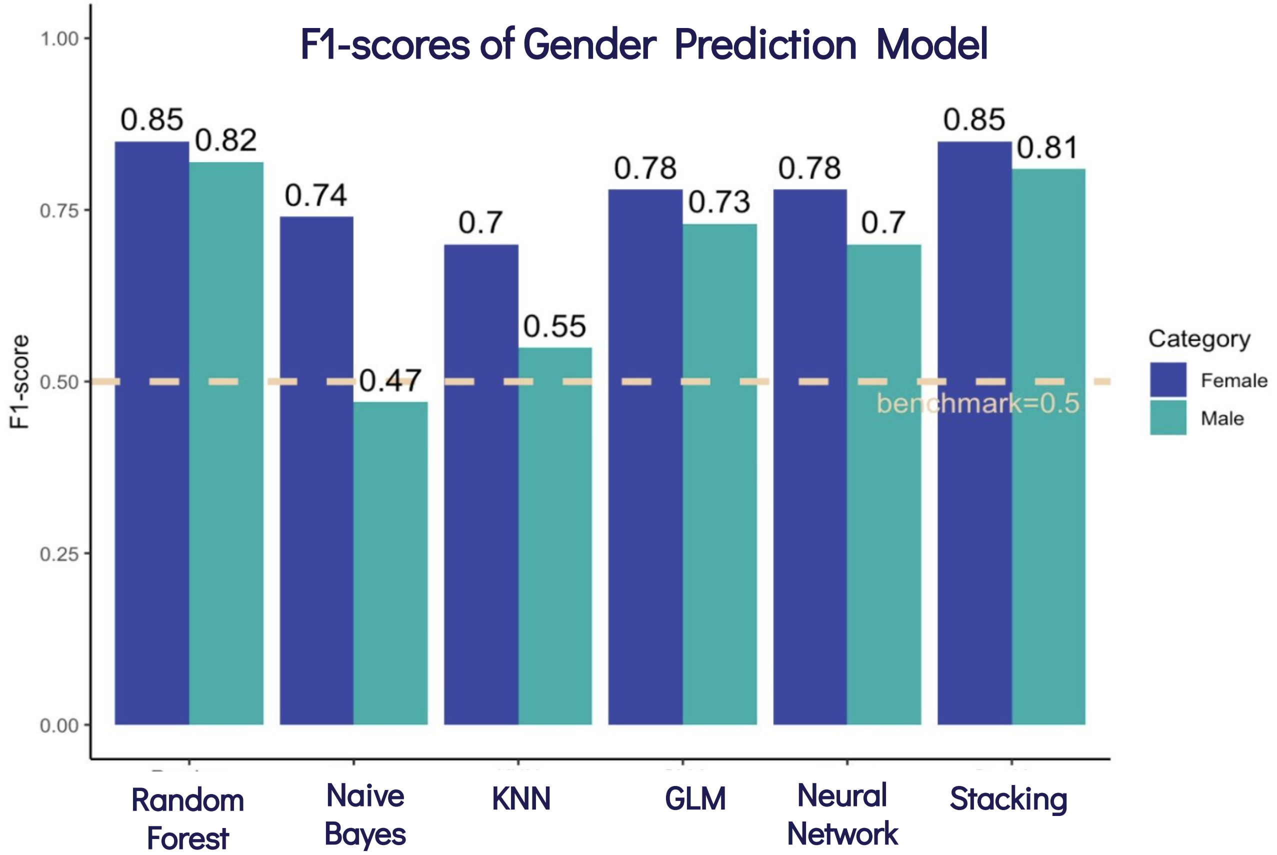
Part 3

Performance Evaluation



Gender Prediction Models' Performance

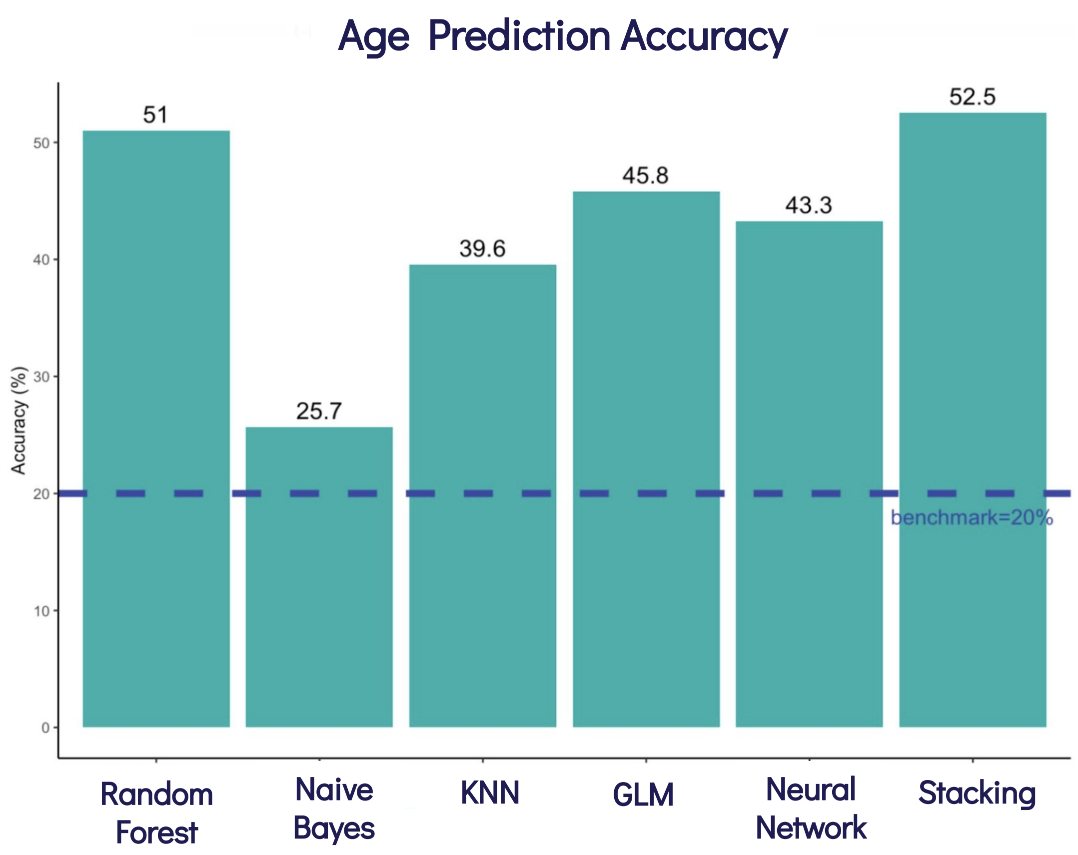


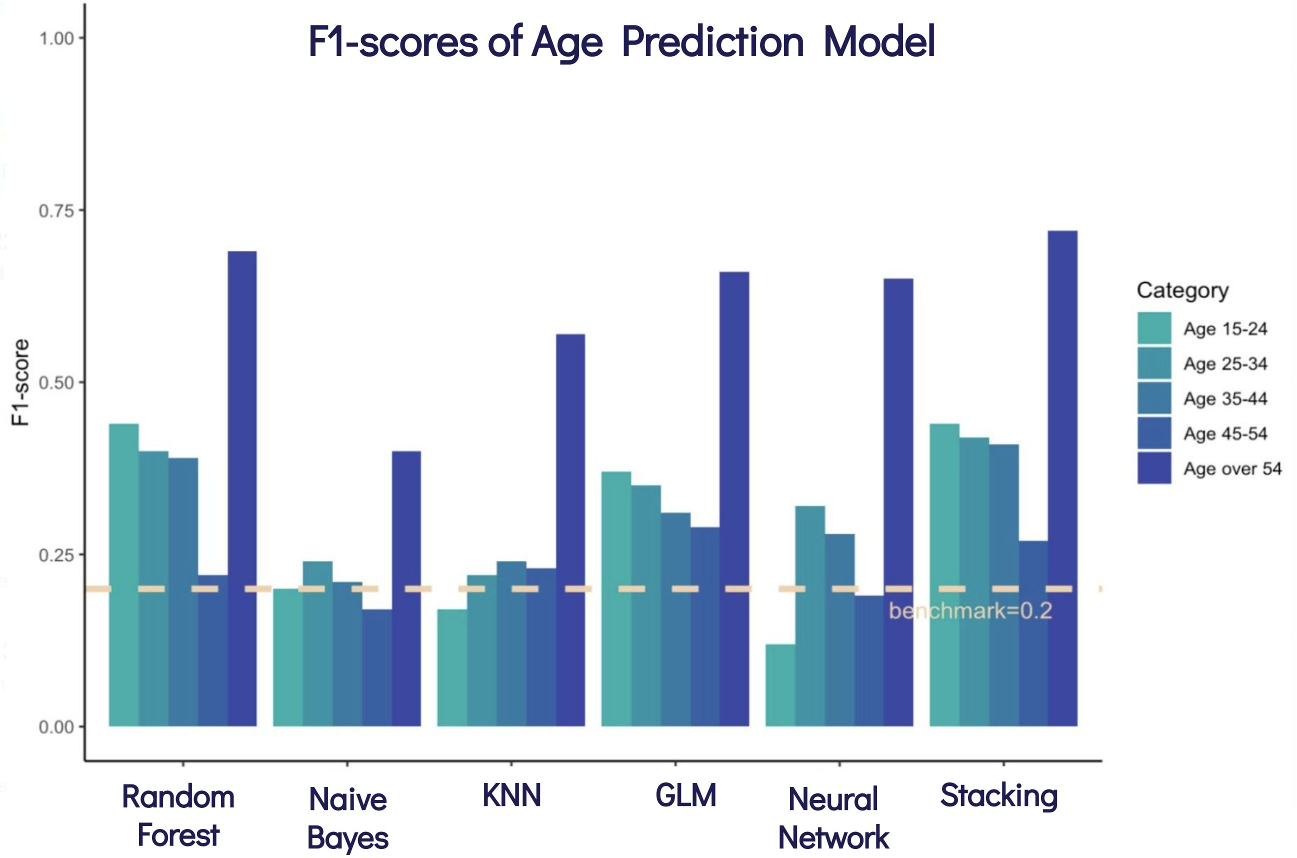


Random Forest performs the best in terms of both accuracy and F1-scores.

Naive Bayes is invalidated because of drastically different F1-scores in its categories.

Age Prediction Models' Performance





Stacking performs the best in terms of both accuracy and f1 scores.

All models have **varying performances** across F1-scores for different age groups.

Runtime Performance - Complex Models Require Higher Runtime

	Random Forest	Naive Bayes	KNN	GLM	Neural Network	Stacking
Gender	1m37s	115ms	18.2s	383ms	1m31s	1m23s
Age	8m24s	209ms	16s	1.5s	2m50s	4m38s

Best Model Overall

Considering both Gender and Age prediction,

Stacking

performs best overall.

Advantages and Limitations of the Stacking model

Advantages

Advantage 1

High accuracy

82.9% accuracy under the cross-val train/test setup is hard to come by in real life settings.

Advantage 2

Low chance of overfitting

Stacking is an ensemble algorithm

Advantage 3

High practicality

Omitted other demographic features (e.g. social status) when predicting gender/age.

Limitations

Limitation 1

Data Representativeness

We only had data for December, which does not represent all the months.

Limitation 2

Black box

We cannot easily interpret the decision process behind the Stacking Model.

Limitation 3

Large volume of features

We have in total hundreds of variables, which cannot be filled in easily with real life data.

Stacking stands as an effective model.

Changes after Generalizing Stacking with Features Selection

~700 features	→	**30** features
82.9% accuracy of Stacking for Gender prediction		**78.3%** accuracy of Stacking for Gender prediction
1m23s execution time		**27s** execution time

Part 4

Applications

Applications - How Our Findings and Models Fit into Reality

SEM Expense Planning

➢ **For Search Engine Marketing (SEM), which keywords should the company spend more on?**

Search Engine Marketing Expense Planning:

- Select top keywords from word importance ranking
- Plan more budget on top keywords

Male

Female

Applications - How Our Findings and Models Fit into Reality

Personalized Online Experience

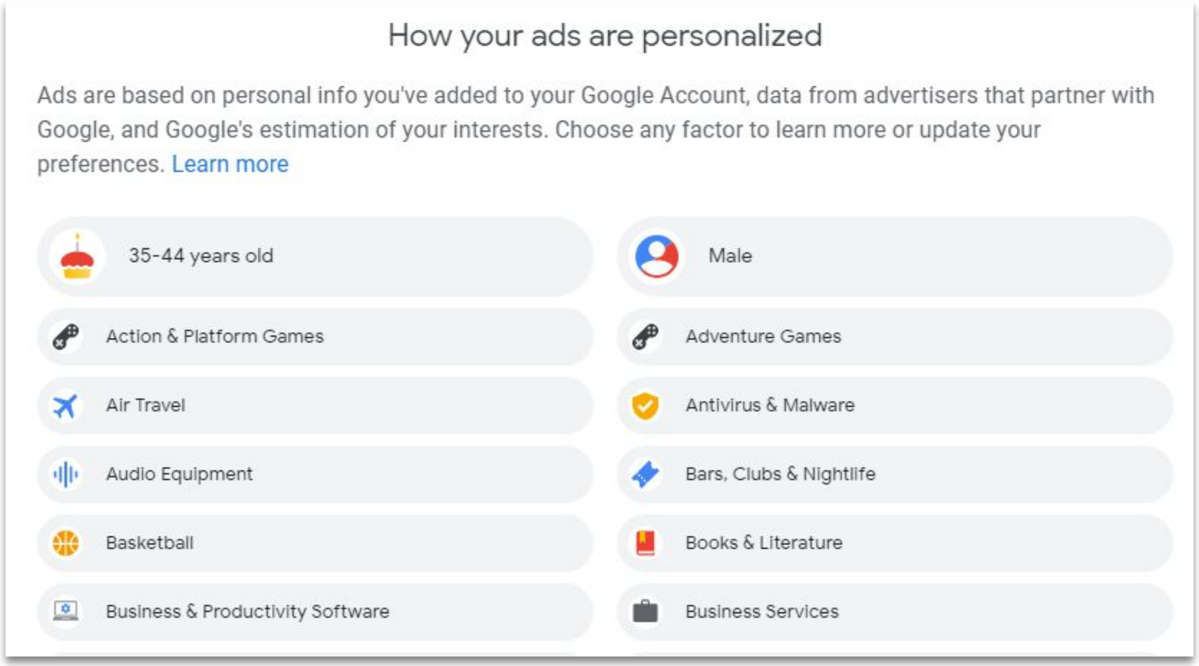
➢ **What can we do with our best model?**

Personalized Marketing:

- Embed our best model to understand the profile of web visitors
- Create more targeted advertising campaigns

Personalized User Experience/User Interface:

- Understand audience
- Design interface to be more visually appealing to target audience



Demographic Prediction by Google

Thank you!