

# Szymon Socha

[linkedin/szymon-socha](#) | [github/szymonsocha](#) | [szymonsocha.github.io](#)

## SUMMARY

Data Scientist/Analyst at Flying Wild Hog. Data Science and Business Analytics final year student at the University of Warsaw. Uses statistical modelling and machine learning algorithms to find new patterns and relationships in data to drive business value. Passionate about using data to solve real-world problems. Avid traveler.

## EDUCATION

### MSc. Data Science and Business Analytics

Warsaw, PL | 2021 - present

UNIVERSITY OF WARSAW

**Coursework:** Machine Learning I, Machine Learning II, Unsupervised Learning, Advanced Econometrics, Applied Finance, Statistics, Python, Advanced Programming in R, Big Data Analytics, Algorithms for Data Science

### BSc. Computer Science and Econometrics

Warsaw, PL | 2018 - 2021

UNIVERSITY OF WARSAW

GPA: 4.5/5

**Thesis:** "The influence of COVID-19 pandemic on the music streaming consumption." An econometric OLS model constructed and tested with R programming language. Python programming language used for data manipulation

**Coursework:** Machine Learning, Econometrics, Linear Algebra, Statistics, Time Series, Probability Calculus, Finance, Python, R, SQL, C++

## WORK EXPERIENCE

### FLYING WILD HOG | DATA SCIENTIST/ANALYST (JUNIOR)

Warsaw, PL | Jan 2022 – present

### PUBLICIS GROUPE | DATA SCIENCE ASSISTANT

Warsaw, PL | Sep 2021 - Nov 2021

- Econometric modelling project.
- Data cleaning and processing with the use of **Python** and **R**.
- Building custom R tools for econometric analysis.

### TESTRONIC | QA TESTER

Warsaw, PL | Feb 2018 – Oct 2018 and Jul 2020 – Oct 2020

- Detect and report bugs (in English), software flaws and design issues on a daily basis (**JIRA**).

## PROJECTS

### TRAFFIC PREDICTION

PYTHON, SKLEARN, RANDOM FOREST, LINEAR REGRESSION, KNN, SVM, RMD

Traffic prediction using multiple machine learning algorithms (Linear, Lasso, Ridge, and Elastic Net Regression, Random Forest, KNN regressorm, SVM). Data manipulation and feature engineering.

### DRUG USE CLASSIFICATION

PYTHON, KNN, LOGISTIC REGRESSION, SVC, RMD

Drug use prediction using multiple machine learning algorithms (Random Forest, k-Nearest Neighbors Classifier, Logistic Regression, Support Vector Classifier). Data manipulation and feature engineering. Balancing data with SMOTE+Tomek. Cross Validation used for parameter tuning.

### STUDENT DROPOUT PREDICTION

R, LOGIT, PROBIT, HYPOTHESIS TESTING

Prediction of student dropout using Logit and Probit models. Hypothesis testing to find the best model. Comparing Logit and Probit models performance. Interpreting the results and making conclusions.

## SKILLS

**Tech stack:** R (tidyverse, dplyr, ggplot2, Rmd, Shiny), Python (sklearn, pandas, numpy), SQL (BigQuery, KQL), Git

**Expertise:** Machine Learning and Econometric data modelling, clusterization, dimension reduction, statistics, hypothesis testing

**Soft skills:** Problem-solving, Stress management, Decision-making, Time management, Communication, Attention to detail

I agree to the processing of personal data provided in this document for realising the recruitment process pursuant to the Personal Data Protection Act of 10 May 2018 (Journal of Laws 2018, item 1000) and in agreement with Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).