# Hallmarks of Online Privacy Minders

By: Burak Basogul

**Abstract**
The potential client(s) for the results of this analysis are software makers who have, internet browsers, VPN's, tracker controllers, privacy ROM's in their product portfolios along with hardware makers who bring privacy into focus in their products.

**Design**

The target group was set from individuals who answered the "How closely, if at all, do you follow news about privacy issues?" question as "very closely". This group made 11% of the survey population, about 470 individuals. (imbalanced dataset!) To try and identify any behavioral patterns of the target group, three machine learning algorithms were employed, using the same training and testing data. Once models were fitted and scored, F1, Recall and Precision scores were used to compare model performances.

**Data**

The original dataset as provided by Pew Research contains 4,272 rows and 143 columns with mostly categorical data. After performing EDA, the column numbers were reduced to 87 and the remaining data was all categorical. Thereafter, the target data was converted to binary type which increased the column numbers to 355.

**Algorithms**

- Feature Engineering

1. All data was converted to binary type using dummy variables.
2. Optimum threshold adjustment made for Logistic Regression model
3. Class weight changed for Random Forest model.

- Models

1. Bernoulli Naïve Bayes
2. Logistic Regression w/Decreased Regularization (Model of choice)
3. Random Forest

- Model Evaluation and Selection

The data set was split into 70/30 train / test portions and the same sets of split train and test data were used to train and score each model. Models were test scored only after all model tuning had been completed on train data. Models were mainly compared using F1, Recall and Precision scores. Eventual Accuracy, F1 and Precision scores were closely grouped however Logistic Regression model's Recall rate was the only one above 50% at 0.581 which was the deciding factor in the end.

**Tools**

- IBM SPSS for data import
- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib, Seaborn and Plot Tree for plotting

**Communication**

- Project presentation slides