**Outline**

The data was taken from the KDD Cup 2009, which is a marketing database from the French Telecom company Orange. The challenge is to predict the propensity of customers to switch provider (churn). This is a classification problem on anonymized, unbalanced data with lots of missing values, and there is also a particular interest in finding a measure that reflects the economic impact of ML models.

**Dataset**

Anonymized data with 190 numeric features and 40 categorical features. Data is unbalanced and contains lots of missing values.

**What was done:**

1. Statistical analysis and visualization
2. Selection of performance score
3. Data preprocessing
4. Baseline modeling (Regression, Random Forest, Gradient Boosting, kNN, Bayes)
5. Feature selection
6. Model parameters fitting via cross validation
7. Model adequacy check on a test set
8. Analysis of economic impact

**Preprocessing**

* OneHotEncoding
* Hashing
* Encoding with densities
* Normalization/ scaling
* Filling missing values
* Rebalancing

**Performance Score**

AUC ROC was used as performance score, as it reflects TPR and FPR. TPR is important for understanding how many churn users are identified and FPR is related to the economical impact (retention is costly).

**Baseline modeling:**

Several algorithms were used: Lasso and Ridge Regression, Random Forest, Gradient Boosting, kNN, Naïve Bayes Classifier. Best result was achieved on Gradient Boosting.

**Feature selection**

Correlation with target, Lasso Regression feature selection.

**Model fitting**

Model was fit with 5-fold cross validation using AUC ROC as a performance measure. Baseline model was Gradient Boosting, parameters were categorical features encoding, filling missing values, sample balancing, features selection as well as Gradient Boosting model parameters that were optimized via Grid Search.

**Best Model**

Best preprocessing tactic was to encode categorical features with densities, fill missing values with an outlier value, delete features with >70% of missing values.

Gradient Boosting parameters:

* loss: ‘deviance’
* learning\_rate: 0.1
* n\_estimators: 100
* max\_depth: 3

**Further Research**

* Feature engineering
* Using of model ensemble
* Manage outliers in numerical features