Report

1. Introduction.

The e-commerce site have lots of new visitors every day. Only around 1% of them complete a purchase in the first 7 days since signup. During my work I was focused on two main problems:

- predict the revenue of a new visitor in the first month based on first week's behavior,
- identify the group of visitor that generates equal / not equal first week and first month purchase based on first week's behavior.

2. Used technologies.

- PostgreSQL,
- Python,
- Scikit,
- PySpark,
- MLLib.

3. Description.

At the beginning I spent some time to get some theoretical knowledge about Machine Learning. It was my first serious encounter with this subject. I finished ML course on the coursera.org platform (https://www.coursera.org/learn/machine-learning). The next part was a meeting with our mentors in the Jagiellonian University campus. Our team worked on the preparation of data and transfer data to the database. We had some problems because data were inconsistent. Moreover we prepared a basic set of features.

The next step was to use Scikit to predict the revenue of a new visitor in the first month based on first week's behavior. I was working with lots of models: linear regression, ridge, lasso etc. Furthermore I was trying imporve my set of features using features selection, cross validation, Kmeans decision trees and other clustering methods. It was very important to us because the our data are very specific. For example 99% of people buy nothing in the first week. 0.4% of people buy something in the first month but nothing in the first week. Also I had a few problems because the set of data is really big for one machine.

The second problem (identify the group of visitor that generates equal / not equal first week and first month purchase based on first week's behavior) was a typical binary classification problem. During my work I was using SVMWithSGD, logistic regression with SGD, logistic regression with LBGS, decision trees, random forests and gradient boosted trees. I was trying decrease test error. Next I was focused on Receiver Operating Characteristic and Area Under the Curve. It was better metric for our set of data because we had more than 99% true results.

4. Some stats.

- 99% of people buy nothing in the first week,
- 0.4% of people buy something in the first month but nothing in the first week,
- AVG of the first week is $1.89, AVG of the first month is $2.9,
- 0.48% of people have not equal first_week and first_month.

5. Summary.

This project was a really big chance for me to get knowledge about Machine Learning. Work atmosphere was great and we could always ask for help our mentors. The biggest problems for me were lack of intuition in Machine Learning and the specific composition of our database. I hope that I will have lots of chance to work with ML in the future.