Stack Overflow Active Contributor Churn and LCV Prediction – MVP Proposal

Songlin Qing

Description: Based on Stack Overflow user contribution history data until Jan 1 2017, predict if the contributor will stay active on Jan 1 2018. For the contributors who remain active, predict their lifetime customer value (LCV).

Data source: https://cloud.google.com/bigquery/public-data/stackoverflow

Data description: Complete history records of questions, answers, and comments posted on Stack Overflow, with timestamps.

Definitions:

Contributor: users who carry out contribution activities (post questions, answers, or comments). Edit activities, though technically also considered as contribution, do not have complete record in this dataset thus excluded.

Churned: users who does not have any contribution activity for a consecutive n days (n to be determined based on activity statistics) are considered as churned.

Customer Value: Stack Overflow has a complex reputation system that rewards users for their contributions and reflect their credibility. I will use a simpler equation to reconstruct the users' reputation by counting their contribution activity since beginning of SO until the measurement date.

Reputation = upvotes on question * 5 +
upvotes on answer * 10 +
accepted answer * 15

Challenges:

- Dealing with large volume of data on BigQuery platform, which I am unfamiliar with
- Going back in time to reconstruct the user reputation at a certain timestamp
- Incorporate the industry best practice (e.g. definition of churn duration, how long in future should we predict etc ..) in this project

Approach:

1. Subset the dataset by question tag (#python) to carve out a small portion of data for prototype building;
2. Explore the data and define churn duration;
3. Perform user segmentation and build classification algorithms;
4. Build LCV regression models;
5. Model selection and performance evaluation.