CSE 523: Machine Learning

Group 17 - Hardly Humans

Weekly Project Report - 9

Quora Insincere Questions Classification

Name	Enrolment Number
Malav Doshi	AU1940017
Parth Shah	AU1940065
Sanya Zaveri	AU1920064
Mihir Pathak	AU1920138

1) Tasks Performed in the week.

- Looking at why we need to balance the dataset.
- Balancing the dataset using different approaches.

2) Outcomes of the tasks performed.

- Rebalanced the dataset using:

  1) Discarding the majority:

  Data has 1.3 million entries of which 93% are sincere questions.

  Since the model would be trained better on a balanced dataset, we remove the excess of sincere questions so now it matches the 7% insincere questions.

  Drawback: Too much data is being discarded.

  2) Oversampling

  The 7% data of insincere questions would be oversampled by duplication to match the 93% of sincere questions.

3) SMOTE (Synthetic Minority Oversampling Technique)

Also an oversampling method, but here the samples are not simply duplicated but they are generated using k-nearest neighbors.

4) Ensemble

Just like random forests, we split the data in multiple sets and train the model multiple times in permutations with different sets. This would result in us getting multiple models for each permutation. From here onwards we would simply choose the best model based on majority vote.

3) Tasks to be performed in the upcoming week.

- Implementing different models like SVM, Random Forest and LDA.
- Looking at performance measures of the different models implemented.