Project Proposal

Junyuan Ke, Shuoyuan Yang, Mingyang Sun

EMORY
UNIVERSITY

Motivation and objective:

For the Final Data Mining project, our group has chosen to take part in the Santander Customer Satisfaction challenge on Kaggle. Santander is a national banking company whose principal market is the northeast of United States. It offers financial services and product including retail banking, mortgages, credit card, insurance, and etc. As a banking corporation, customer satisfaction is not only crucial to Santander's success in maximizing profit but also necessary in building long-term customer loyalty. However, collecting customer feedback could be troublesome to even the best customer service, as not all customers voice their concerns. Moreover, unsatisfied customers are more likely to cancel their service without leaving a complaint, making it harder for banks to collect sufficient information for analyzation. Therefore, an accurate model that precisely predicts customer satisfactions based on service experience would enable Santander to take precaution and more likely to enhance a bad customer's experience in advance.

For this competition, we were given a dataset of banking experience from Santander, with all the features anonymized. Our goal is to predict the probability whether a customer is satisfied or unsatisfied based on his or her banking experience. As this is a supervised learning problem, the target attribute of the training data is labeled with whether the customers are satisfied or unsatisfied with their experience at Santander. Since this is a categorical(binary) classification problem and scores are measured through area under the ROC curve, our team is planning to approach this problem by constructing different classification models as well as trying combined models on predicting customer satisfaction. First, we are going to test out simple individual models like decision tree, Bayes' network, SVM, KNN, random forest and logistic regression. After evaluating our basic models with ROC curve, we plan to utilize ensemble methods to increase accuracy by combining models and assigning them with different weights. By learning different models, bagging and boosting methods can also enhance overall accuracy.

Our goal is to establish a compound model in predicting customer satisfaction for Santander. Not only can this model be used for Santander's future customer satisfaction prediction, it can also shed light on similar problems in the tertiary industry.

Related work/methods:

This is a classical classification problem: whether the customer is happy or not? This dataset has over 200 attributes, while some attributes contain over 80% missing values (0). Data cleaning could help us remove comparatively useless attributes and noisy data. Different attribute selection techniques-such as the function between a specific feature and the class label-provide us with a guide. The basic methods for this problem are (1) Decision Tree Induction and (2) naive Bayes Classifier. Improvements on classifiers to get better accuracy include: apply the Chow-Liu Algorithm to construct a Bayesian Belief Network, build a gradient boosted tree based on weights, or construct neural networks (really complex). The evaluation model is crucial in this case: there is just a small group of dissatisfied customers. Therefore, the most naive evaluation methods will probably suggest that our classification methods work well even if we classify all customers as "happy". To avoid this implicit error, pre-processing on the training dataset, like oversampling or undersampling, could be applied to make our model more sensitive to classify unhappy customers.

Proposed work:

The first step of our work would be data preprocessing. Since the datasets contains some invalid values, we need to first figure out those values and edit them. For some attributes with too many invalid values, we may consider the deleting it, and in general, an attribute selection process might be needed to better filter out the redundant data. After that, we can start working on classification methods on those selected data. Bayes classifier can be incorporated to generate the decision tree from the training data. Based on the simplistic version of the decision tree, we can then optimize the classification method. The potential improvements are discussed above in the methods section. The most amount of work would be spent on the optimization part in order to get a better accuracy. The last step is to evaluate the classification model. As discussed above, the evaluation method also needs to be optimized so that the actual quality of this classification would be not be affected by the small proportion of unhappy customers.

Evaluation:

Our dataset is from Santander Bank. The class label is "happy" (TARGET=0) or "unhappy" (TARGET=1). This dataset has 369 attributes and 76,021 tuples. By observation, this dataset has two explicit features: 1) it has a large number of attributes but most attributes have lots of missing values: they are numerical attributes but over 80% of values are 0; 2) the number of class label "unhappy" (TARGET=1) is few compared to that of label "happy". To overcome this shortcoming, our evaluation model is based on the combination of weighted F-beta score and ROC curve: the model with the highest F-beta score and the largest area under the ROC curve will be chosen. As in reality Santander Bank is concerned more about reaching unhappy customers as early as possible, the beta which makes the model comparatively more sensitive to correctly detect unhappy customers (higher recall) will be picked.

Plan of action:

0309 - First Meeting


- Get familiar with ipynb and python, as well as working with git.
- Scrap through kernels on Kaggle discussion forum for ideas and insights.
- Work on project proposal.

0322 - Project Proposal


- Work on data preprocessing: Attribute selection, data reduction, handle noise, missing data.
- Attempt simple classification models via sklearn and other packages.
- Group meeting on April 4th.

0405 - First Submission


- Optimize classification methods and setup evaluation model on the finished classifier.
- Work on accuracy improvement; finish adjustment on evaluation model.
- Prepare slides and other relevant work for the presentation, Group meeting on Apr 16th.

0417 - Presentation

-Work on presentation feedbacks.
-Make final submissions; finish final report.

0504 - Final Report