Membership plan signup response model

Audible Data Science
11/15/2018

Agenda

- Problem Statement & Scope
- Data manipulation & Modeling
- Insights/Recommendations
- Appendix

Problem Statement and Scope

Problem Statement:

To predict users who sign up for our membership plan

Scope:

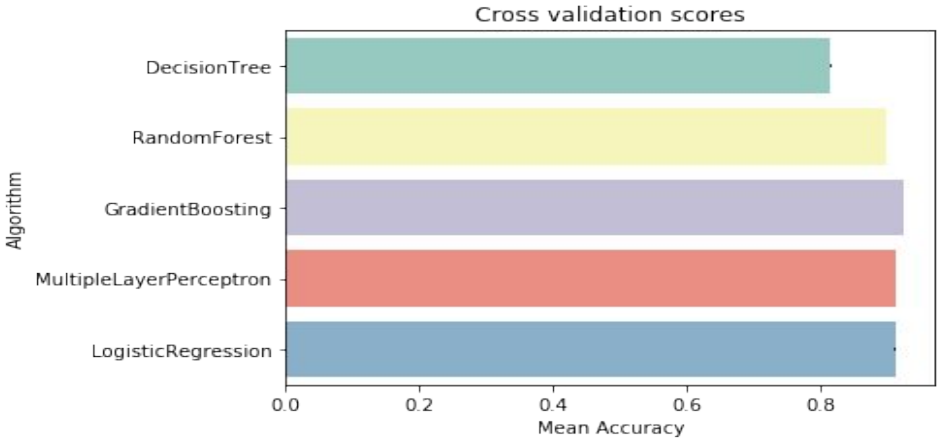
- Given total target population :~351K customer
- Independent variables: 171 features
- Dependent variable : 1, if a user has signed up for the plan,otherwise 0
- ~91K users have missing response indicator,so dropped those rows
- Final data size consists of ~259K users and ~78K responders, having a response rate of 30.09%

Data Manipulation and Modeling

- Missing value imputation
- Defining train and validation data
- Building baseline models using various machine learning techniques
- Hyper-parameter tuning for selected techniques
- Performance evaluation and model results

Data Manipulation and Modeling

- Missing value imputation is done by replacing nulls with -999
- Train and validation data was created by performing random 70-30 split of the entire data set
- Baseline models are created using 3 fold cross validation and default hyper parameter settings



Counts	#Contacts	#Response	Response Rate
Train	~181K	~54K	~30%
Validation	~77K	~23k	~30%

Hyper-parameter tuning

Grid search based iterative hyper-parameter tuning is performed for RF and GBM models

Final Parameter metrics for Random Forest:

Mean accuracy increased from 0.8984 to 0.9183

max_depth: 11,
max_features: 19,
min_samples_leaf: 11,
min_samples_split: 200,
n_estimators: 200

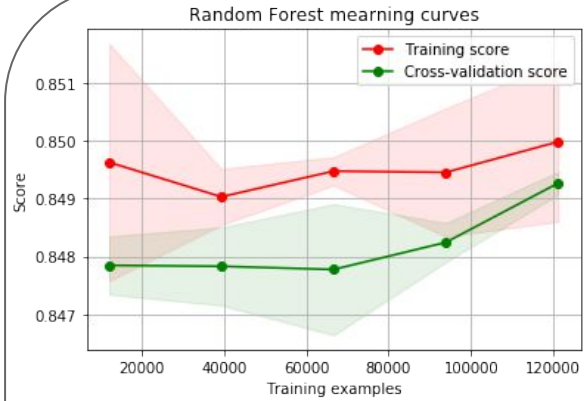
Final Parameter metrics for Gradient Boosting Machine:

Mean accuracy increased from 0.9239 to 0.9270

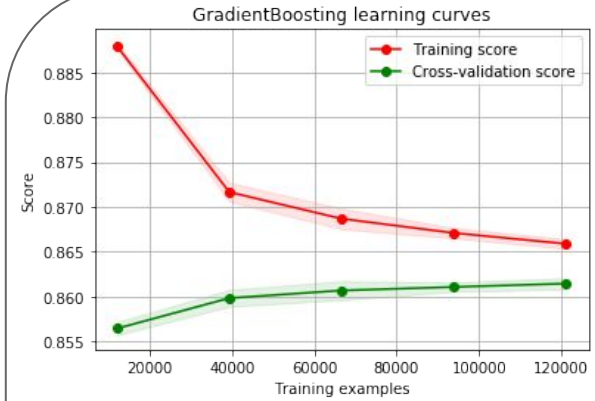
learning_rate: 0.01,
max_depth: 5,
max_features: 11,
n_estimators: 3000,
subsample: 0.85

Performance evaluation

As the no of training examples increase both training and cross validation curves tend to come closer and so both the trees seem to better generalize the prediction

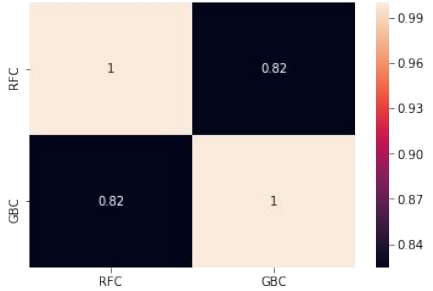


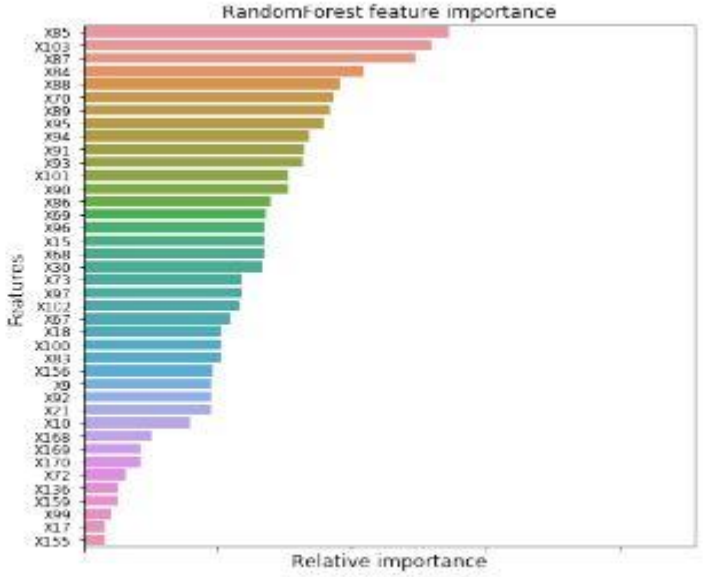
roc_auc: 0.767333660615
gini: 0.53466732123
F1 score: 0.84

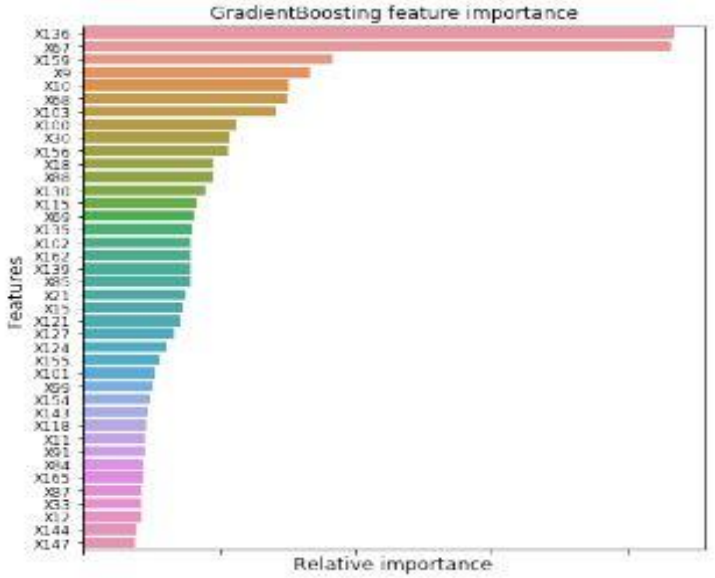


roc_auc: 0.807658442172
gini: 0.615316884345
F1 score: 0.86

Prediction seems to be quite similar for both classifiers







Recommendations

Model can be testing in three areas:

1) Targeting right customers:

- Test top deciles vs random, especially useful in campaigns involving huge budget eg. direct mail campaigns

2) Targeting customers at right time:

- Targeting users when a change in the probability score (spikes in the curve) is observed on a daily,weekly or monthly basis depending on the nature of the features

3) Feature importance:

- Top features can be used in email creatives (subject lines,placements,designs) or as talking points for customer care representatives (in phone channel)

Thank you!

Appendix

Here is the link to code base