Classification

Long Zhao

21 October, 2022

Background

Lots of non-tech companies are quite conservative in adopting analytic tools. This phenomenon is even more severe in highly regulated industries, like finance and healthcare. For example, in one fortune 500 bank[1], there is a team consisting of about 20 Ph.D.s focusing on evaluating model risks. All models need this team's approval to be adopted widely across the company. There are cases that several employees used three months to develop a model but fail to obtain permission. In general, the models with good interpretability is easier to pass the review. Unfortunately, those models tend to perform worse than the hard-to-explain ones. This means that the managers might be unsatisfied with the performance, and one could not submit the model in the first place. In this project, you will try to pass this review with a model with decent performance. That is to say, you need to balance **between interpretability and performance.** More specifically, you could choose one of the following.

1. Use logistic regression or **simple** classification tree to predict. You need to convince me that the performance of your model is good enough. Namely, they are not far away from the hard-to-explain models.
2. Use hard-to-explain models to predict. You need to analyze the models thoroughly and convince me about the following two claims.
    - There is no overfitting.
    - The most important features are intuitive.

Introduction

An insurance company wants to sell vehicle insurance to its policyholders of health insurance. Its data science team wants to target potential buyers using analytics. Here is a glance at data.

id	Gender	Age	Region_Code	Age.1	VehicleDamage	Premium	Vintage	Response	Mystery
319654	Male	22	46	< 1 Year	Yes	29719	79	0	M3
298528	Female	40	28	1-2 Year	Yes	29645	99	0	M20
120858	Male	25	28	1-2 Year	Yes	35996	266	1	M3
234326	Female	25	29	< 1 Year	No	56899	25	0	M20
372307	Female	70	28	1-2 Year	No	47452	189	0	M17
212306	Male	41	46	1-2 Year	Yes	19851	219	0	M17

Here is the description of the data.

[1] I need to protect the privacy of my friends who work in this bank.

Variable	Definition
id	Unique ID for the customer
Gender	Gender of the customer
Age	Age of the customer
Region_Code	Unique code for the region of the customer
Age	Age of the Vehicle
Damage	1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
Premium	The amount customer needs to pay as premium in the year
Vintage	Number of Days, Customer has been associated with the company
Response	1 : Customer is interested, 0 : Customer is not interested
Mystery	20 different categories

The target variable is Response, and covariates are the other variables.

Note

This project is based on Health Insurance Cross Sell Prediction from Kaggle. I have modified data to make the problem simpler.

Techincal Requirements - Part 1

In this part, you do not use Mystery!!!

First of all, the performance measure adopted is the AUC of the **validation (could be CV)** set. That is to say, you should compare models based on it. Secondly, different choices of prediction models have different requirements as shown below.

- Logistic Regression.
  - Why you think its performance is good enough?
  - Are the coefficients reasonable?
- Classification Tree.
  - What are the hyperparameters of the tree? For example, the minimum number of observations in a leaf or the maximum depth of the tree.
  - Do you obtain a nice explanation based on the classification tree?
  - Could you plot the tree? If the tree is too big, you might answer you can not plot the tree.
- Gradient Boost and Other Models.
  - **Random Forest does not count because too many students used it before.**
  - What are the hyperparameters?
  - Which features are the most important? Why they make intuitive sense? Please generate the importance plot.
  - For the important features, are there any categorical variables? If so, should one be concerned about the number of categories it has?
  - Why do you want to choose this hard-to-explain model?
  - **You need to do a little tuning of hyperparameters: the final model should have a better validation or CV performance than the classification tree.**
- No matter which model you choose, please compare the model performance on the validation set (or CV) and the test set.
  - Is the test set performance worse than the validation set? If so, why? If not, why?

Techincal Requirements - Part 2

In this part, you use Mystery and other covariates!!! You will compare the feature importance of two different encodings of Mystery. The method you will use is random forest.

- For Python users, you will use integer encoding and **dummy** encoding.
  - Integer encoding: M1 < M2 < . . . < M20
- For R users, you will use integer encoding and **factor** encoding.
  - Integer encoding: M1 < M2 < . . . < M20

Please provide some explanations of what you observe in the feature importance. Hint for Python users, what is the impact of random selection of covariates?

Discussion

1. Do you think AUC is a good performance measure? Why? If not, could you propose a better one?
2. For the chosen model, what is your probability threshold (one of 1%, 10%, 20%, 50%) if the benefit structure is
   - Promote to an interested customer + 10
   - Miss an interested customer - 10
   - Promote to an uninterested customer - 2
   - Each promotion - 1
   - Please explain why?
3. Answer question 2 again with the new benefit structure
   - Promote to an interested customer + 100
   - Miss an interested customer - 100
   - Promote to an uninterested customer - 2
   - Each promotion - 1
   - Please explain why?
4. Do you think combining logistic regression and the hard-to-explain model is a good idea? Why?

Hint: you might need to code to answer question 2 and 3.

Report Requirements

- Providing reasonable predictions result only gives you 70% of grade.
  - The other 30% come from the good explanation and discussion questions.
  - Playing safe will not be enough! Think, think, think.
- Pages: 6 pages (excluding cover & reasonable font) without penalty
  - You do not need to report all things in detail. Just the ones that fit your storyline.

Deliverables.

- Report + Code: separately to Canvas.

FAQ

sklearn.linear_model.LogisticRegression by default have L2 penalty. You shall one-hot encoding all categories! Namely, you shall set drop = None.

1. If I choose a classification tree, do I need answers for other methods?
   - No, you do not. Your target is to have the chosen model to pass the evaluation. Thus, there is no need to justify other models.
2. Do I need to tune hyperparameters?
   - Logistic regression: there is no hyperparameter to tune.
   - Classification tree: tune hyperparameters such that it is better than the random guess.
   - Hard-to-explain model: tune hyperparameters such that it is better than the classification tree.
3. What is a nice explanation of a classification tree?
   - If the tree is simple, one could get a nice interpretation of the tree. For example, in the diamonds example in class, the classification tree only has 3 leaves.
4. Do I have to plot the ROC?
   - No, just a value is enough.
5. Should I deal with the imbalance of data?
   - No, you do not need to. If you want to explore this direction, you shall compare the models with treatment to the imbalance issue to the models without. That is to say, I view your treatment which is changing the weights of data as another hyperparmeter.

Reference.