Cancellation Rate Prediction of Auto Insurance Policy

Shuaimin Kang, Yubing Yao

University of Massachusetts Amherst

January 24, 2018

Data Exploration

Prediction Models

Summary

Discussion

Data Exploration

Data:

- 7578 records, 16 features, 105 NAs, 25 cancel$== -1$, 40 age $<$ len.at.res, 8 age $>$ 100.

High Cancellation population:

- Low credit, Virginia, 2014(mainly caused by Virginia and Washington), Single + young + Online + New customer + Higher premium, Coverage type B + Tenant, Tenant + (n.adults > 3), Male + high n.adults, House + White color

Interactive visualization details:

- `https://visualplot1.herokuapp.com/`

Prediction Models

A set of candidate models with good prediction performance through cross validation of AUC score in training dataset.

Base Models:

- **Linear models**: bagging on logistic regression with Lasso method, linear SVC

Prediction Models

A set of candidate models with good prediction performance through cross validation of AUC score in training dataset.

Base Models:

- **Linear models**: bagging on logistic regression with Lasso method, linear SVC
- **Tree based ensemble models**: Extra Tree classifier with Lasso, Extreme gradient boosting classifier (Xgboost)

Prediction Models

A set of candidate models with good prediction performance through cross validation of AUC score in training dataset.

Base Models:

- **Linear models**: bagging on logistic regression with Lasso method, linear SVC
- **Tree based ensemble models**: Extra Tree classifier with Lasso, Extreme gradient boosting classifier (Xgboost)
- **Other models**: Neural Network, k-nearest neighbor (knn)

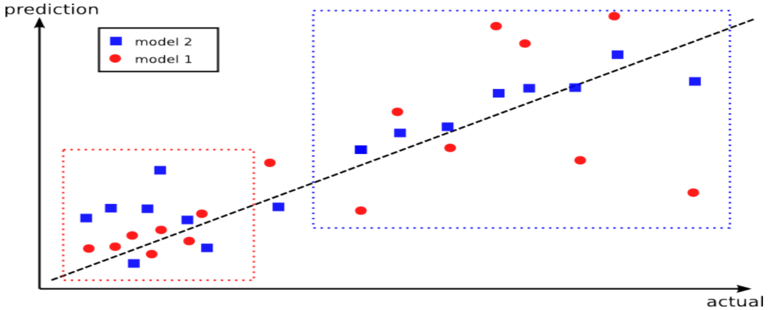
Important features in prediction models

We identify most important features for predicting cancellation rate based on two tree based models-**Decision Tree** and **Extreme Gradient Boosting (Xgboost)** and regression coefficients of the features in **logistic regression** with Lasso method.

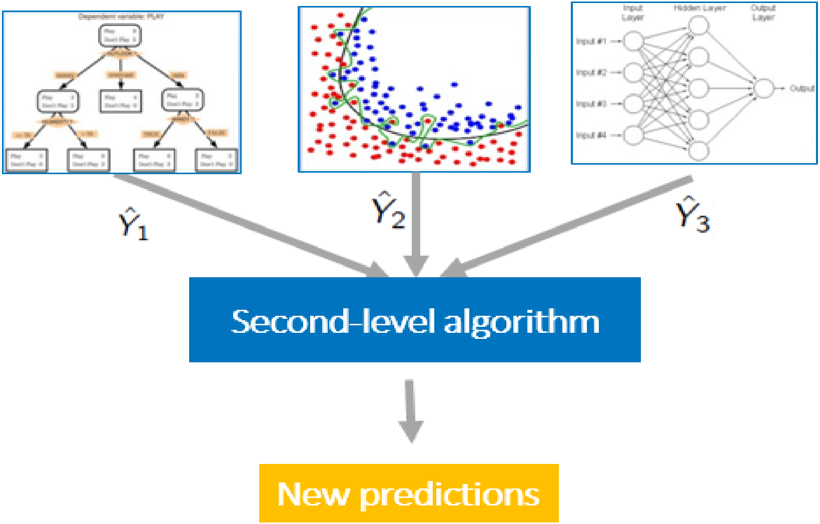
`https://dhn6zvu0qoinq.cloudfront.net/`

Stacking models-Motivation

Data is quite heterogeneous and one single model can only capture partial information in the data. Combining multiple diverse models through stacking we could improve prediction accuracy.



Stacking multiple base models-an illustration



$\hat{Y}_1$ $\hat{Y}_2$ $\hat{Y}_3$

Second-level algorithm

New predictions

Second level algorithm in stacking models

1. We use logistic regression as second level algorithm combining predicted probabilities of 6 base prediction models.

Second level algorithm in stacking models

1. We use logistic regression as second level algorithm combining predicted probabilities of 6 base prediction models.

2. We further reduce variation and over-fitting of stacking model by bagging the second level logistic regression.

Second level algorithm in stacking models

1. We use logistic regression as second level algorithm combining predicted probabilities of 6 base prediction models.

2. We further reduce variation and over-fitting of stacking model by bagging the second level logistic regression.

3. Generally speaking stacking models has better prediction performance using more diverse base models.

Summary

1. We did data pre-processing, like imputation for NAs, and we built an app for interactive visualization.

Summary

1. We did data pre-processing, like imputation for NAs, and we built an app for interactive visualization.

2. We used lasso logistic regression and decision tree models to extract important features and interactions, such as credit level, claim indication, Virginia, young and Online buyers.

Summary

1. We did data pre-processing, like imputation for NAs, and we built an app for interactive visualization.

2. We used lasso logistic regression and decision tree models to extract important features and interactions, such as credit level, claim indication, Virginia, young and Online buyers.

3. We applied advanced modeling techniques, such as bagging, boosting and stacking, to reach a robust prediction.

Discussion

1. Why Virginia high cancellation, Pennsylvania low?

Discussion

1. Why Virginia high cancellation, Pennsylvania low?
2. How to attract more young, new customers through Online?

Discussion

1. Why Virginia high cancellation, Pennsylvania low?
2. How to attract more young, new customers through Online?
3. Explore deeper on “perfect” profile policyholders. Why they even cancel?

Discussion

1. Why Virginia high cancellation, Pennsylvania low?
2. How to attract more young, new customers through Online?
3. Explore deeper on "perfect" profile policyholders. Why they even cancel?
4. Tracking and modeling policyholders' features and cancellation behaviors dynamically.

Discussion

1. Why Virginia high cancellation, Pennsylvania low?
2. How to attract more young, new customers through Online?
3. Explore deeper on “perfect” profile policyholders. Why they even cancel?
4. Tracking and modeling policyholders’ features and cancellation behaviors dynamically.
5. Direct/indirect causal effect? Confounder?

Thank You!