HUIZHE (SUNNY) ZHU

GLM, AUTOML, AND LOAN DEFAULT DETECTION

TAKE-AWAYS AND INSIGHTS - WEEK 10

PROJECT STUCTURE

1. Load and explore data

2. Feature selection

3. EDA on selected features

4. GLM without regularization and with regularization

5. autoML: Fit model + Model evaluation

6. Model evaluation: ROC AUC, the cumulative Lift

1. LOAD AND EXPLORE DATA

- Create Variables_Dictionary to make sense of column names

```
Variables_Dictionary
```

]:

	var	description
3	AP001	YR_AGE
4	AP002	CODE_GENDER
5	AP003	CODE_EDUCATION
6	AP004	LOAN_TERM
7	AP005	DATE_APPLIED
...	...	...
254	PA023	DAYS_BTW_APPLICATION_AND_FIRST_COLLECTION_CALL
259	PA028	AVG_LEN_COLLECTION_OR_HIGH_RISK_CALLS
260	PA029	AVG_LEN_COLLECTION_OR_HIGH_RISK_INBOUND_CALLS
261	PA030	AVG_LEN_COLLECTION_OR_HIGH_RISK_OUTBOUND_CALLS
262	PA031	AVG_LEN_COLLECTION_CALLS

2. FEATURE SELECTION

3 steps:

1.Based on missing values

- Variables 'TD044', 'TD048','TD051','TD054','TD055','TD061','TD062' have more than 79990 missing values and no variable description, we have 80000 rows in our dataset, thus drop these 7 variables

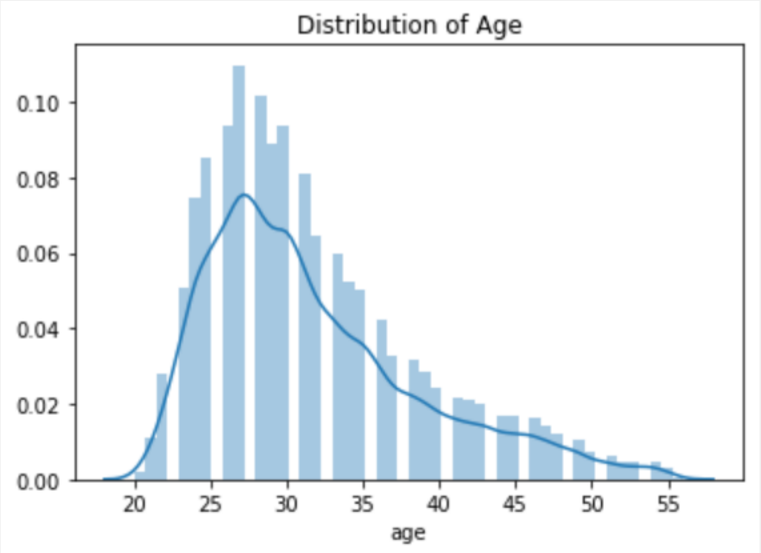
2. Based on correlation

- Remove one of two features that have a correlation higher than 0.9, 20 variables are removed in this step

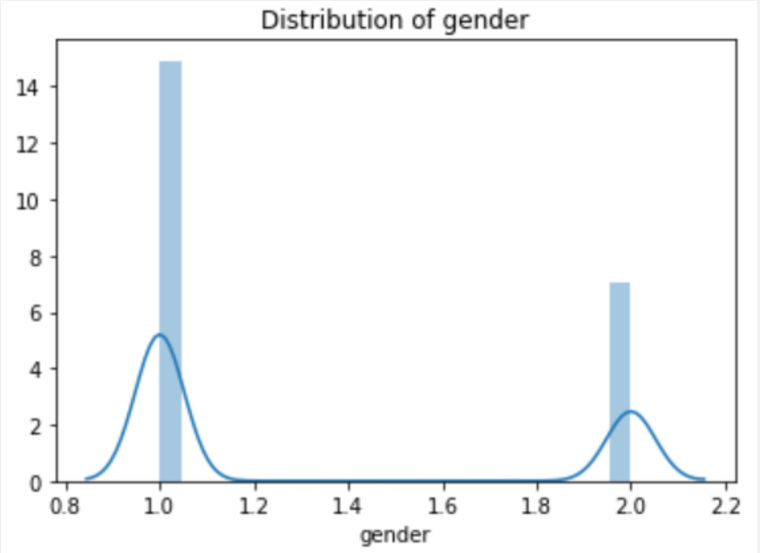
3. Based on decision tree importance

- We only keep variables that are significantly important in random forest model, we keep the top 30 variables

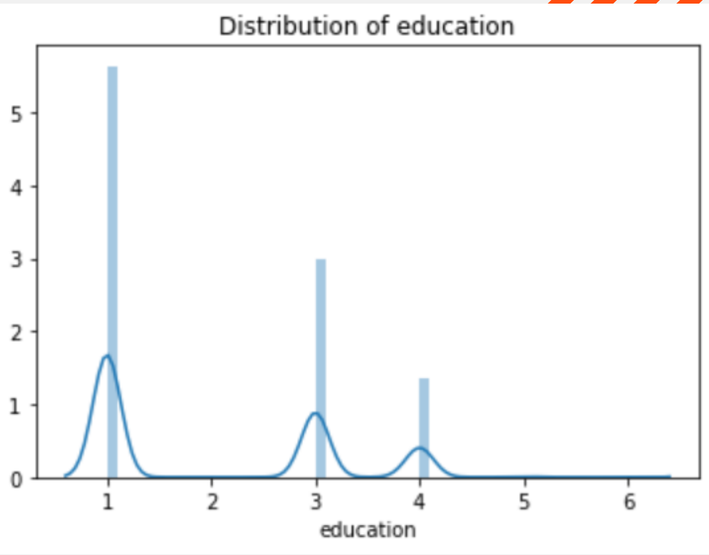
3. EDA ON SELECTED FEATURES



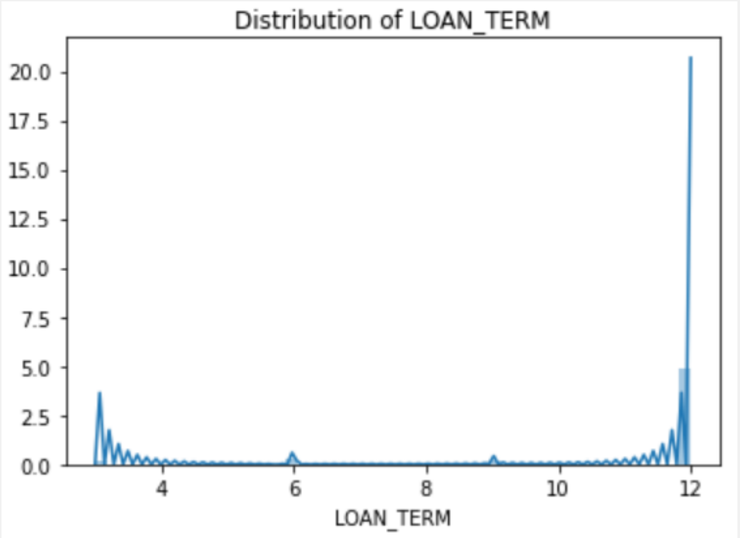
age group: 24-32 is most common



gender imbalance



lower education level



long-term debt application is most common

12 months > 3 months > 6 months > 9 months

4.1 GLM: INTRODUCTION

Definition
Generalized Linear Models (GLM) estimate regression models for outcomes following exponential distributions

Regularization
Put penalty if there are too many coefficient in loss function

Hyper-parameters

- lambda: Specify the regularization strength. Lambda: 0, no regularization; Lambda>0, have regularization
- lambda_search: Specify whether to enable lambda search, starting with lambda max (the smallest λ that drives all coefficients to zero).
- alpha: Specify the regularization distribution between L1 and L2. The default value of alpha is 0 when SOLVER = 'L-BFGS'; otherwise it is 0.5. when alpha = 0, ridge, L2, when alpha = 1, lasso, L1.

4.1 GLM: MODEL PERFORMANCE

Without regularization

- **AUC: 0.688**
- **lift: 2.1**
- **Area under PR curve: 0.3411**

With regularization

- **AUC: 0.688**
- **lift: 2.12**
- **Area under PR curve: 0.3409**

```
glm_v2
```

```
Model Details
=============
H2OGeneralizedLinearEstimator :  Generalized Linear Modeling
Model Key:  GLM_model_python_1606083491961_6


GLM Model: summary
```

	family	link	regularization	lambda_search	number_of_predictors_total	number_of_active_predictors	number_of_iterations
0	binomial	logit	Elastic Net (alpha = 0.5, lambda = 0.001739)	nlambda = 100, lambda.max = 0.1144, lambda.min = 0.001739, lambda....	34	29	62

Summary

- GLM with regularization has higher lift, while GLM without regularization has higher AUC.
- GBM after tuning parameters from last week has the highest AUC.

4.2 AUTOMATIC ML

AutoML

Train all algorithms, rank by their performances, and then choose the best

```
### Leaderboard: compare diff models' performance
aml_v1.leaderboard.head()
```

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
StackedEnsemble_AllModels_AutoML_20201122_175533	0.683585	0.45873	0.33114	0.366927	0.381798	0.14577
StackedEnsemble_BestOfFamily_AutoML_20201122_175533	0.683347	0.458807	0.330956	0.367667	0.381819	0.145786
GBM_grid__1_AutoML_20201122_175533_model_1	0.681183	0.460904	0.329319	0.368685	0.382575	0.146364
GBM_2_AutoML_20201122_175533	0.660199	0.48036	0.302481	0.382003	0.390523	0.152509
DeepLearning_1_AutoML_20201122_175533	0.657893	0.468667	0.302402	0.385891	0.386181	0.149136
GBM_3_AutoML_20201122_175533	0.654973	0.480127	0.297936	0.386504	0.3904	0.152412
GLM_1_AutoML_20201122_175533	0.654276	0.480881	0.293793	0.388628	0.390717	0.152659
GBM_4_AutoML_20201122_175533	0.649236	0.479663	0.292189	0.388119	0.390194	0.152252
GBM_1_AutoML_20201122_175533	0.638687	0.484032	0.28242	0.39221	0.392009	0.153671
GBM_5_AutoML_20201122_175533	0.627272	0.485305	0.274398	0.406149	0.392538	0.154086

Performance

- roc auc: 0.6826
- accumulative lift: 2.06
- PR: 0.3339

Not surprisingly, ensemble model has the best performance

SUMMARY

In summary, the models have similar performance.

If comparing models in this week, GLM without regularization has higher AUC, while GLM with regularization has higher lift, GLM overperforms AutoML

If combine results from last week, GBM after tuning parameters has the highest AUC, GLM with regularization has higher lift

Please drop ideas on why GLM with regularization defeats AutoML, thank you!