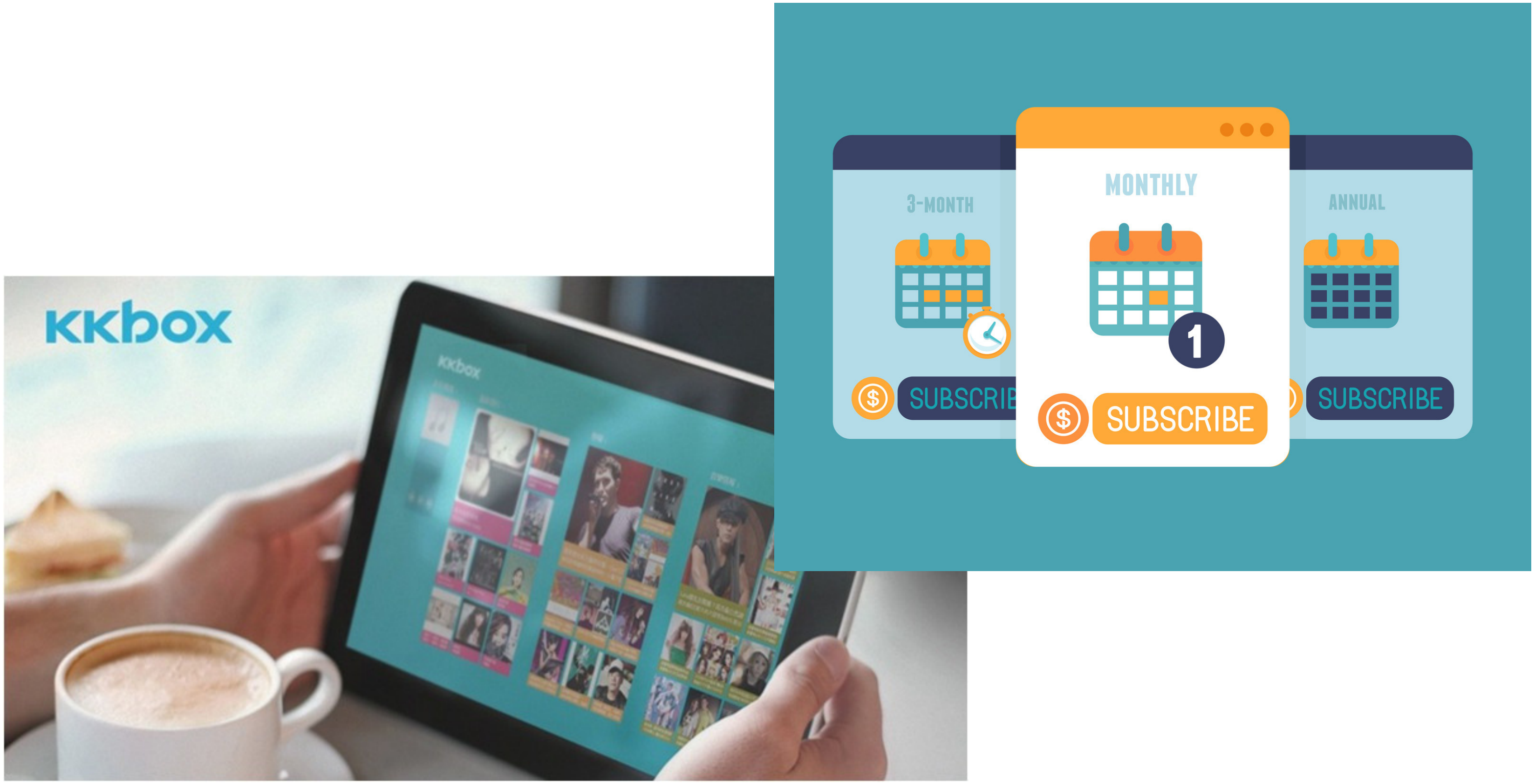
To churn
or not to churn?

The KKBOX
Kaggle Challenge

100%
Python

DATA MINING I PROJECT
TEAM 9
HWS 2017

KKBOX and the meaning of Churn



The DATA

members

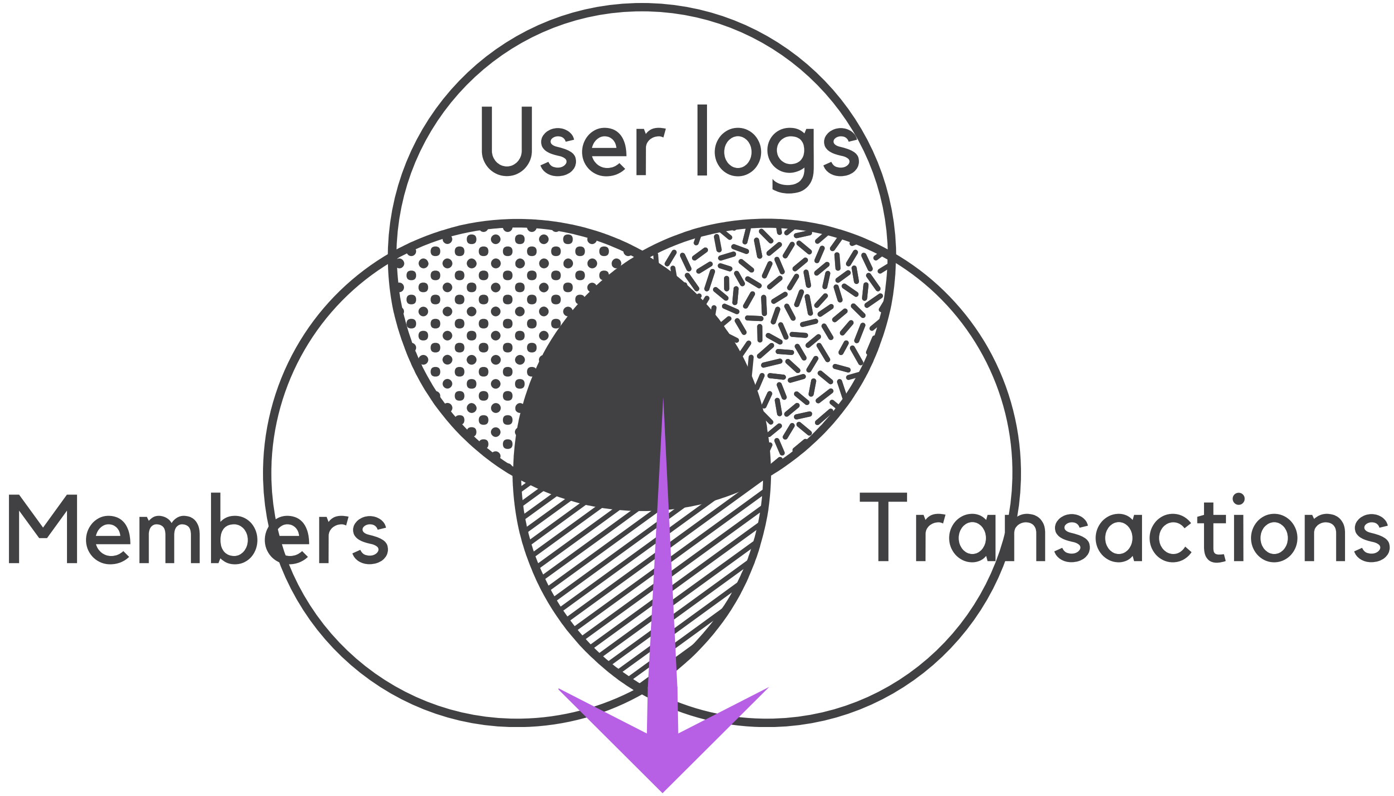
user logs

transactions

train

Data Preparation

Reduce height : Data selection and aggregation



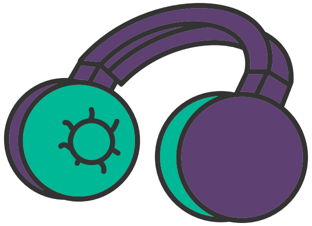
6 months view

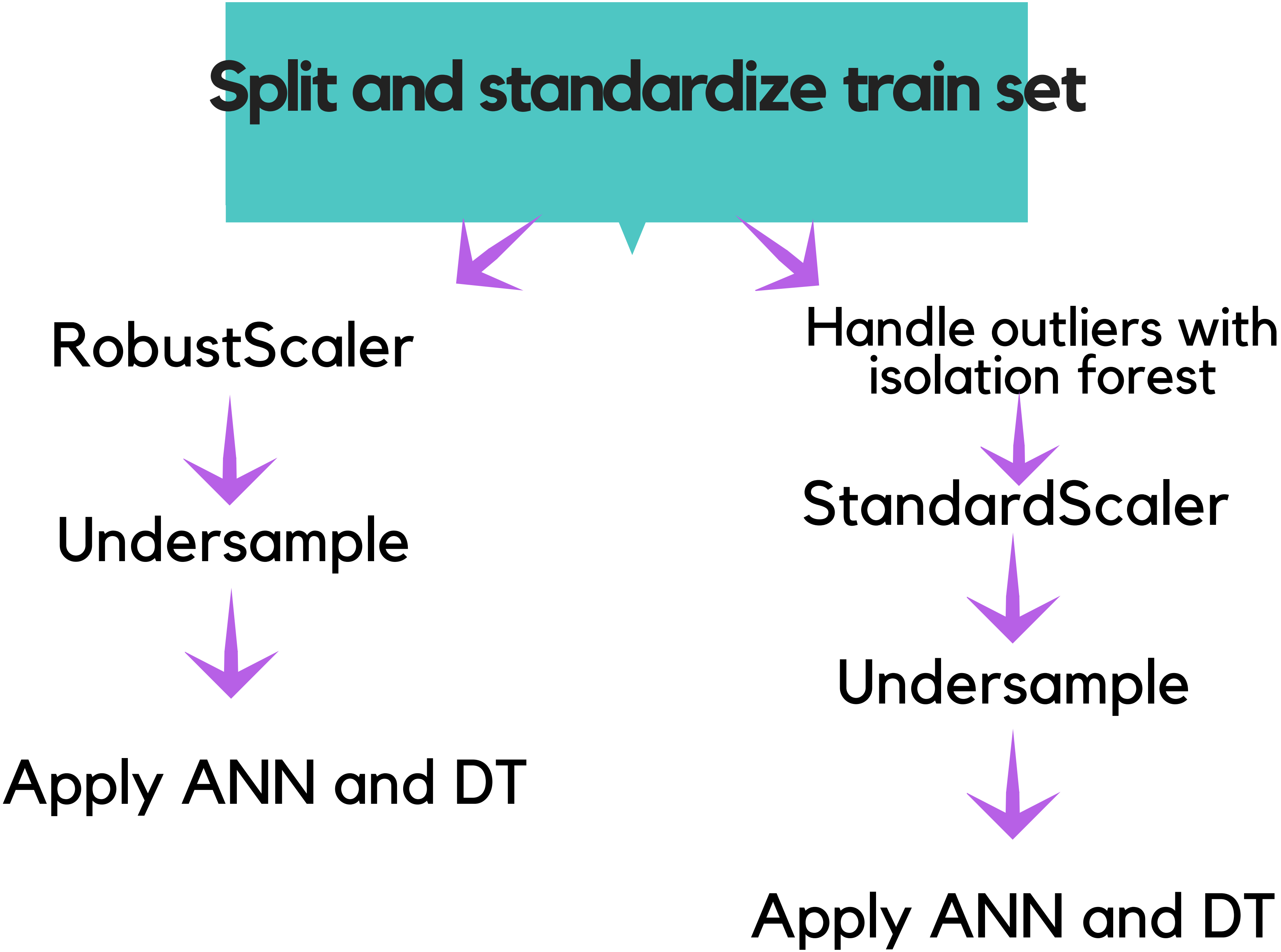
Data Cleaning and Transformation

Generating new attributes:
- total_churn based on transactions
- monthly aggregated usage activities

Removing abnormal values:
- age up to 2000?
- days below 0?

Treating missing and null values:
- categorical: replace with most common
- numeric: replace with mean





Split and standardize train set

RobustScaler

Undersample

Apply ANN and DT

Handle outliers with isolation forest

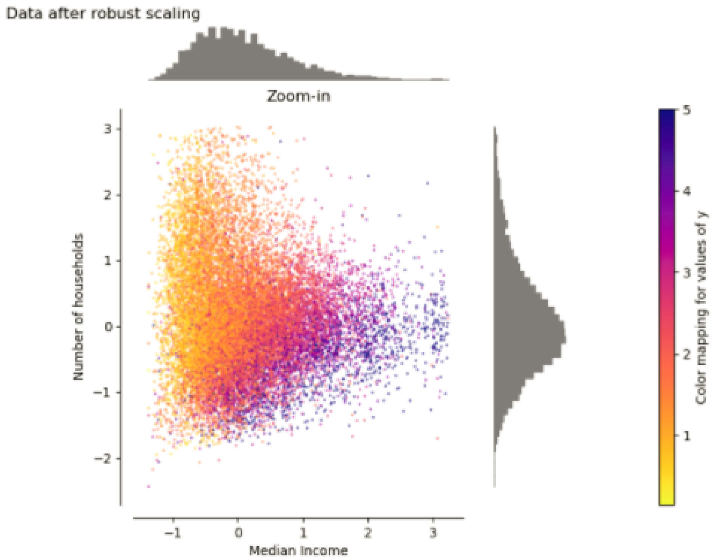
StandardScaler

Undersample

Apply ANN and DT

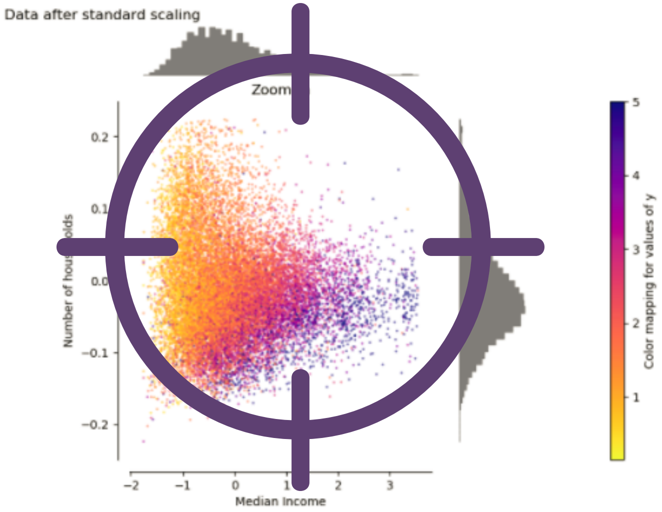
Balance train dataset and evaluate

RobustScaler



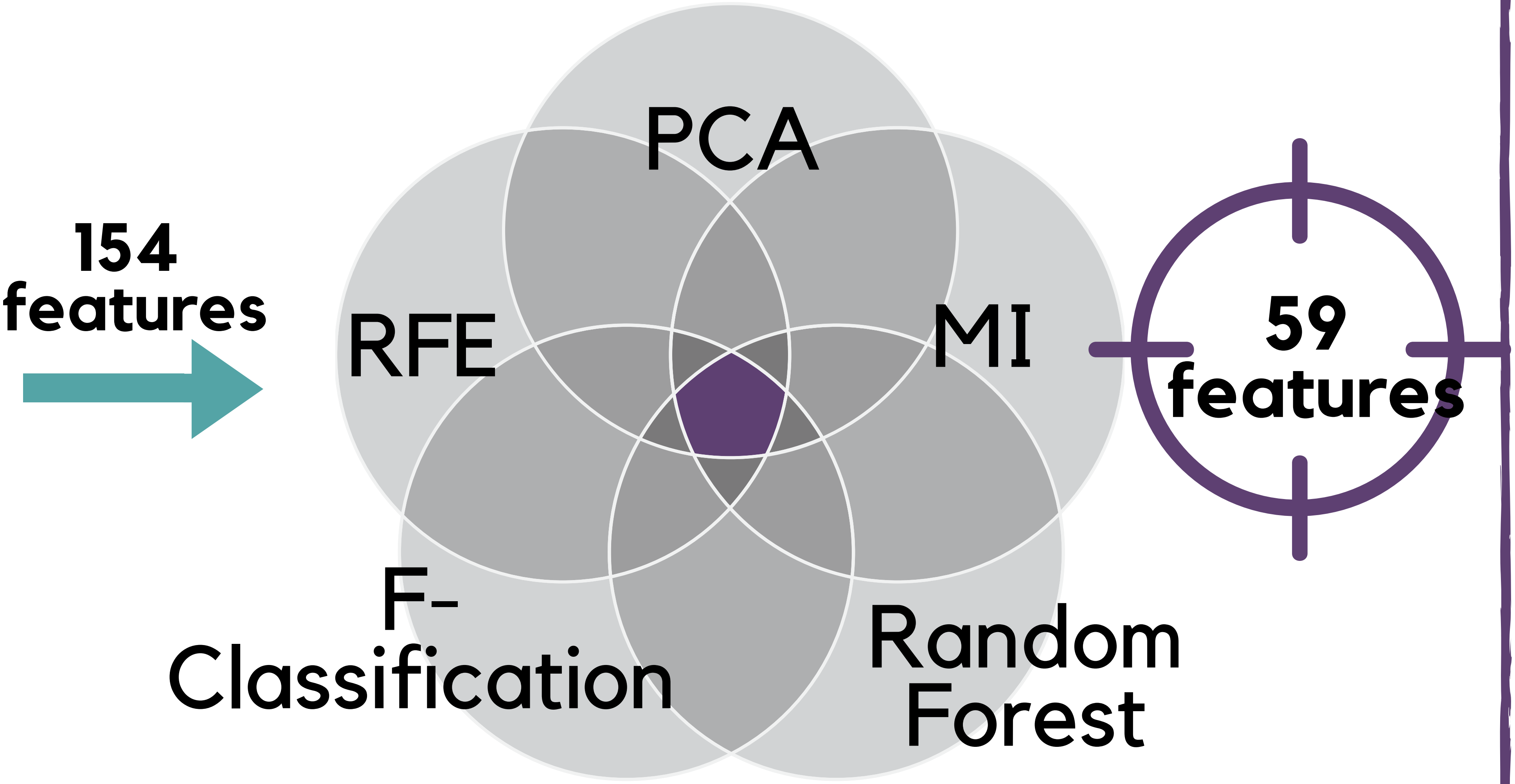
ANN: 90.82
DT: 88.59

StandardScaler



ANN: 91.89
DT: 89.66

Reduce width: Feature selection



154 features

PCA

RFE

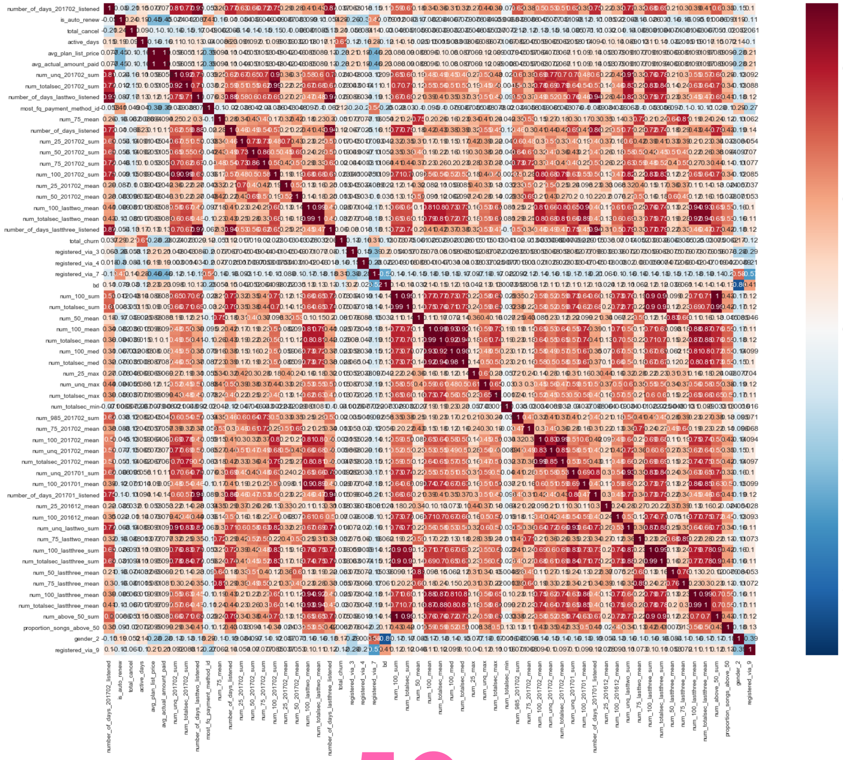
MI

F-Classification

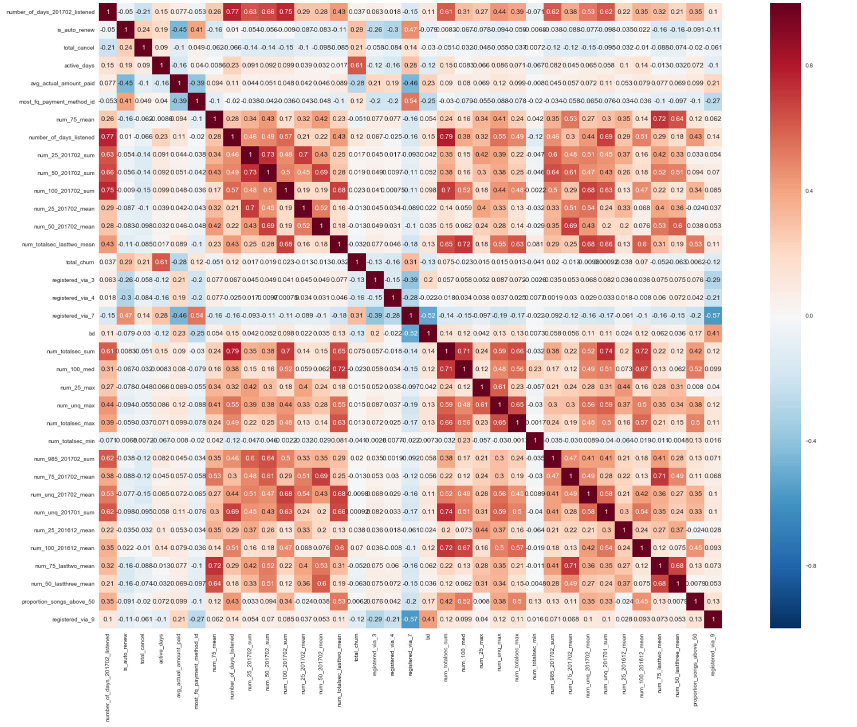
Random Forest

59 features

Reduce width: Correlation Matrix



59



35

Data Mining

52
Models

Logistic Regression
SVM
K-NN
Nearest Centroid
Naive Bayes
Decision Tree
ANN

In Search for the best model...

Approach: start with a baseline model for each algorithm, fine-tune the parameters locally and then compare the chosen best models globally .

- K-NN : optimal number for K?
- DT : combination of parameters, pure nodes removal?
- ANN : optimiser function, features, hidden layers/units?
- Logistic Regression : optimal features subset, C value?
- SVM : kernel, optimal gamma, C ?
- Naive Bayes/ Nearest Centroid : optimal features subset?

Grid Search

Evaluation Metrics

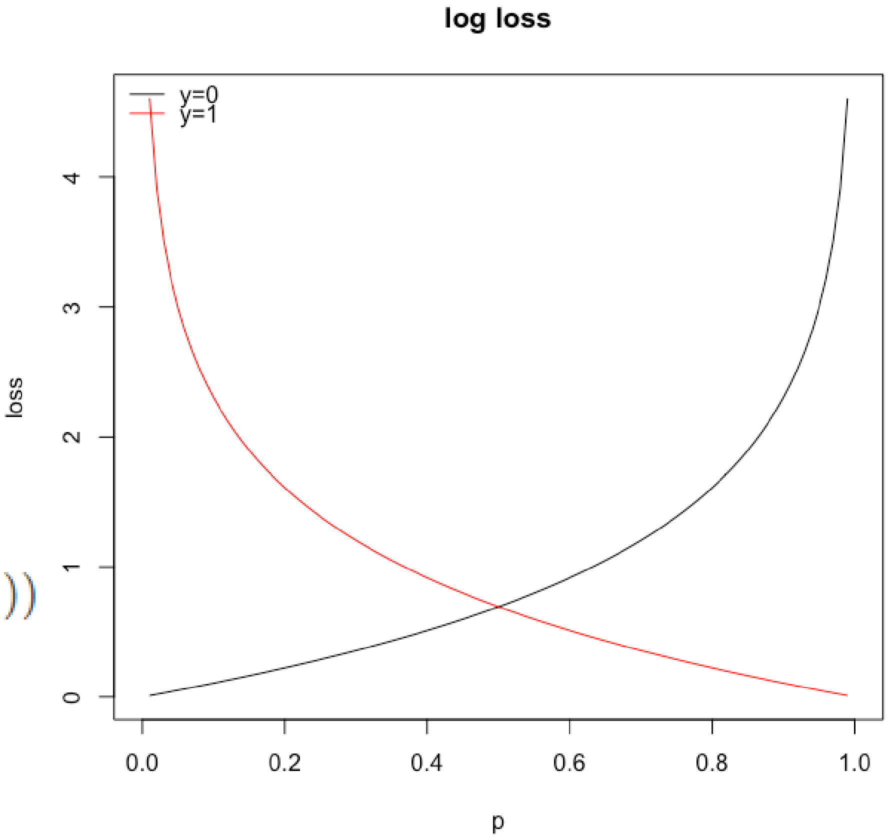
⚠ - Not accuracy: count of correct predictions

- Recall , Precision and F1-score

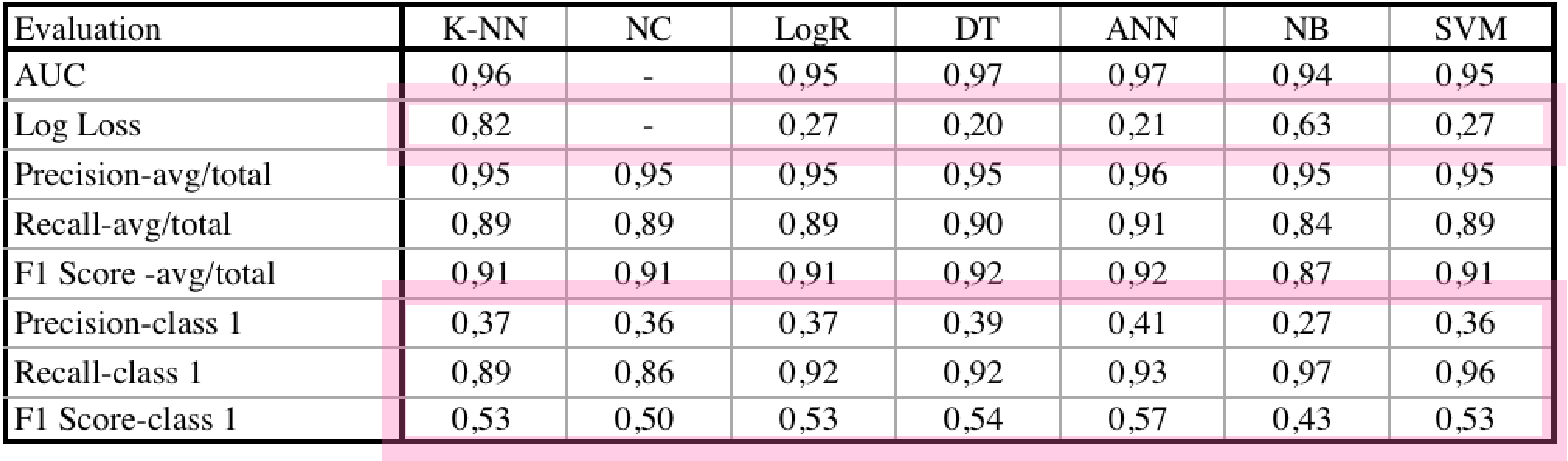
- ROC curve and AUC

- **Log-loss** : account for the uncertainty of your prediction

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

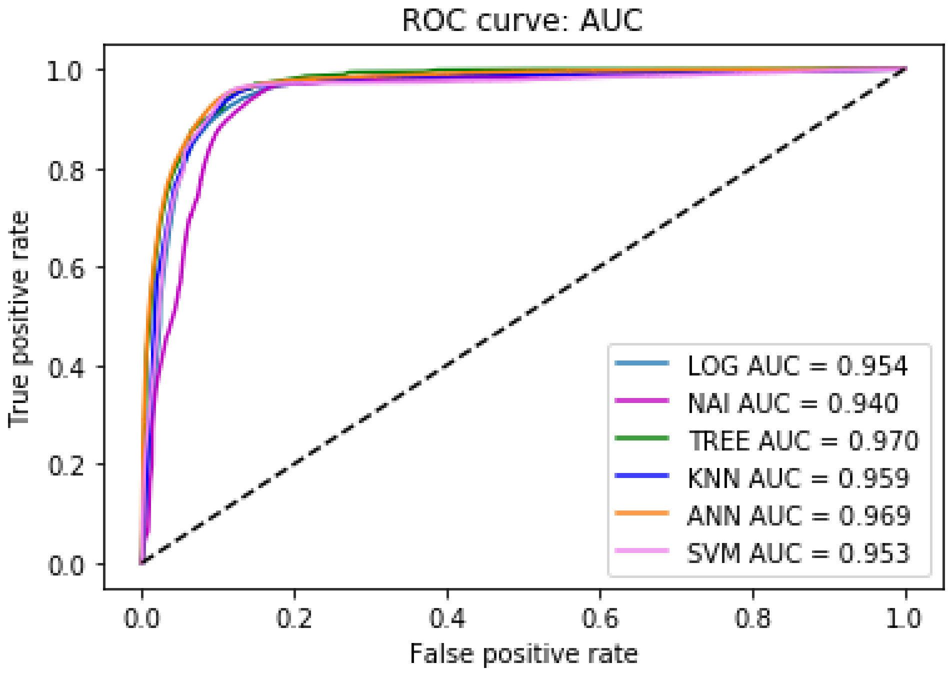


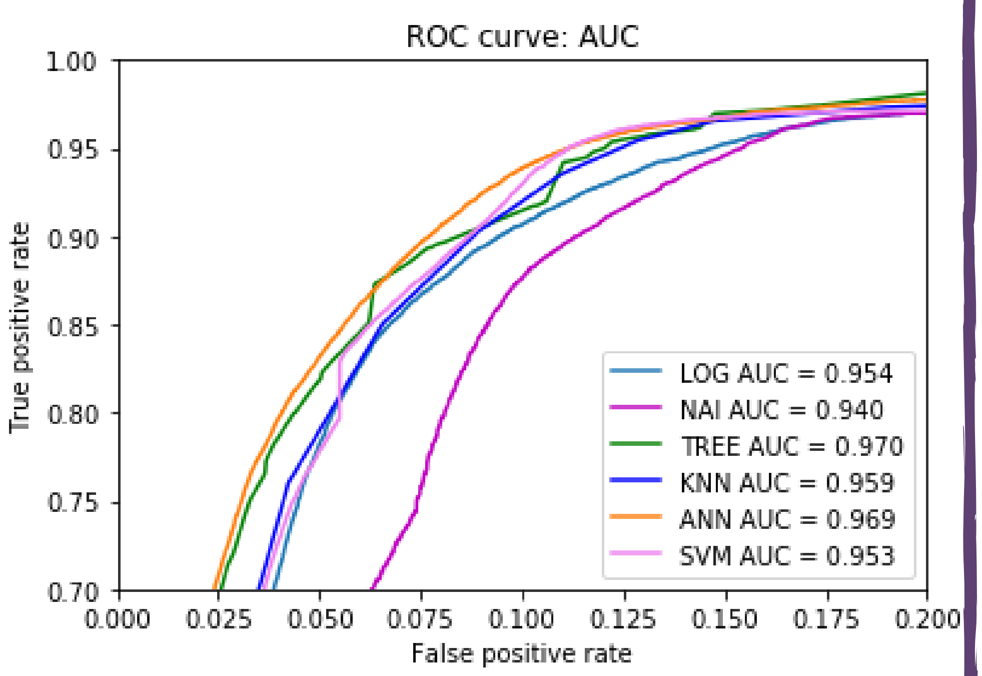
Performance Evaluation



Evaluation	K-NN	NC	LogR	DT	ANN	NB	SVM
AUC	0,96	-	0,95	0,97	0,97	0,94	0,95
Log Loss	0,82	-	0,27	0,20	0,21	0,63	0,27
Precision-avg/total	0,95	0,95	0,95	0,95	0,96	0,95	0,95
Recall-avg/total	0,89	0,89	0,89	0,90	0,91	0,84	0,89
F1 Score -avg/total	0,91	0,91	0,91	0,92	0,92	0,87	0,91
Precision-class 1	0,37	0,36	0,37	0,39	0,41	0,27	0,36
Recall-class 1	0,89	0,86	0,92	0,92	0,93	0,97	0,96
F1 Score-class 1	0,53	0,50	0,53	0,54	0,57	0,43	0,53

ROC curve AUC



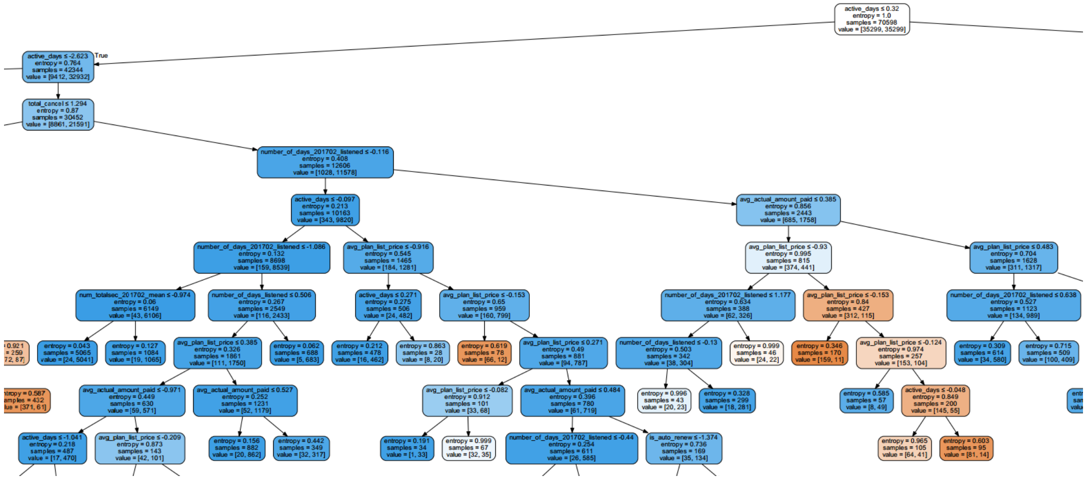


Conclusion and Outlook

#1

# Decision Tree

Optimal parameters:
- criterion='entropy',
- min impurity decrease=0.000015,
- max depth=15,
- max leaf nodes=None,
- min samples leaf=10,
- min samples split=20


active_days ≤ 0.32 entropy = 1.0 samples = 70598 value = [35299, 35299]
True
active_days ≤ -2.623 entropy = 0.764 samples = 42344 value = [9412, 32932]
total_cancel ≤ 1.294 entropy = 0.87 samples = 30452 value = [8861, 21591]
number_of_days_201702_listened ≤ -0.116 entropy = 0.408 samples = 12606 value = [1028, 11578]
active_days ≤ -0.097 entropy = 0.213 samples = 10163 value = [343, 9820]
avg_actual_amount_paid ≤ 0.385 entropy = 0.856 samples = 2443 value = [685, 1758]
number_of_days_201702_listened ≤ -1.086 entropy = 0.132 samples = 8698 value = [159, 8539]
avg_plan_list_price ≤ -0.916 entropy = 0.545 samples = 1465 value = [184, 1281]
avg_plan_list_price ≤ -0.93 entropy = 0.995 samples = 815 value = [374, 441]
avg_plan_list_price ≤ 0.483 entropy = 0.704 samples = 1628 value = [311, 1317]
num_totalsec_201702_mean ≤ -0.974 entropy = 0.06 samples = 6149 value = [43, 6106]
number_of_days_listened ≤ 0.506 entropy = 0.267 samples = 2549 value = [116, 2433]
active_days ≤ 0.271 entropy = 0.275 samples = 506 value = [24, 482]
avg_plan_list_price ≤ -0.153 entropy = 0.65 samples = 959 value = [160, 799]
number_of_days_201702_listened ≤ 1.177 entropy = 0.634 samples = 388 value = [62, 326]
avg_plan_list_price ≤ -0.153 entropy = 0.84 samples = 427 value = [312, 115]
number_of_days_201702_listened ≤ 0.638 entropy = 0.527 samples = 1123 value = [134, 989]
= 0.921 = 259 72, 87]
entropy = 0.043 samples = 5065 value = [24, 5041]
entropy = 0.127 samples = 1084 value = [19, 1065]
avg_plan_list_price ≤ 0.385 entropy = 0.326 samples = 1861 value = [111, 1750]
entropy = 0.062 samples = 688 value = [5, 683]
entropy = 0.212 samples = 478 value = [16, 462]
entropy = 0.863 samples = 28 value = [8, 20]
entropy = 0.619 samples = 78 value = [66, 12]
avg_plan_list_price ≤ 0.271 entropy = 0.49 samples = 881 value = [94, 787]
number_of_days_listened ≤ -0.13 entropy = 0.503 samples = 342 value = [38, 304]
entropy = 0.999 samples = 46 value = [24, 22]
entropy = 0.346 samples = 170 value = [159, 11]
avg_plan_list_price ≤ -0.124 entropy = 0.974 samples = 257 value = [153, 104]
entropy = 0.309 samples = 614 value = [34, 580]
entropy = 0.715 samples = 509 value = [100, 409]
ntropy = 0.587 amples = 432 lue = [371, 61]
avg_actual_amount_paid ≤ -0.971 entropy = 0.449 samples = 630 value = [59, 571]
avg_actual_amount_paid ≤ 0.527 entropy = 0.252 samples = 1231 value = [52, 1179]
avg_plan_list_price ≤ -0.082 entropy = 0.912 samples = 101 value = [33, 68]
avg_actual_amount_paid ≤ 0.484 entropy = 0.396 samples = 780 value = [61, 719]
entropy = 0.996 samples = 43 value = [20, 23]
entropy = 0.328 samples = 299 value = [18, 281]
entropy = 0.585 samples = 57 value = [8, 49]
active_days ≤ -0.048 entropy = 0.849 samples = 200 value = [145, 55]
active_days ≤ -1.041 entropy = 0.218 samples = 487 value = [17, 470]
avg_plan_list_price ≤ -0.209 entropy = 0.873 samples = 143 value = [42, 101]
entropy = 0.156 samples = 882 value = [20, 862]
entropy = 0.442 samples = 349 value = [32, 317]
entropy = 0.191 samples = 34 value = [1, 33]
entropy = 0.999 samples = 67 value = [32, 35]
number_of_days_201702_listened ≤ -0.44 entropy = 0.254 samples = 611 value = [26, 585]
is_auto_renew ≤ -1.374 entropy = 0.736 samples = 169 value = [35, 134]
entropy = 0.965 samples = 105 value = [64, 41]
entropy = 0.603 samples = 95 value = [81, 14]

		Predicted	
		0	1
Actual	0	219,970	25,580
	1	1436	15,857

Discussion

Drawbacks:
- other balancing approaches: oversampling
- scale out to the whole timeframe (3 years)

Outlook:
- hybrid model*

*Lee, Jae, and Jin Lee. "Customer churn prediction by hybrid model." Advanced Data Mining and Applications (2006): 959-966.

Thank You !

Presented by
Team 9:

Simona Doneva
Na Gong
Bengi Koseouglu
Siying Liu
Liam Pang

Questions?

YOU ARE A DATA SCIENTIST
YOU ARE A DATA SCIENTIST
YOU ARE A DATA SCIENTIST
YOU ALL ARE DATA SCIENTISTS!

Decision Tree on train data before standardization and balancing

Optimal parameters:
- criterion='entropy',
- min impurity decrease=0.000015,
- max depth=15,
- max leaf nodes=None,
- min samples leaf=10,
- min samples split=20

		Predicted	
		0	1
Actual	0	243,240	2310
	1	8212	9081

Top Features

- active days
- total churn
- total cancel
- actual payment
- registration method
- total_sec_listened
- total_days_listened

....