2023 Winter 33002 Intro to ML
Individual Project

Customer Churn Prediction in Telecom Company

Lee Kyung ko

Background

- **Customer churn**

: Core business metric for business operation across industries.

Essential to retain existing customers and target profitable customers

- From Machine Learning perspective

: **Binary classification task to predict customer churn**

- Telco Customer Churn data ([Kaggle link]())

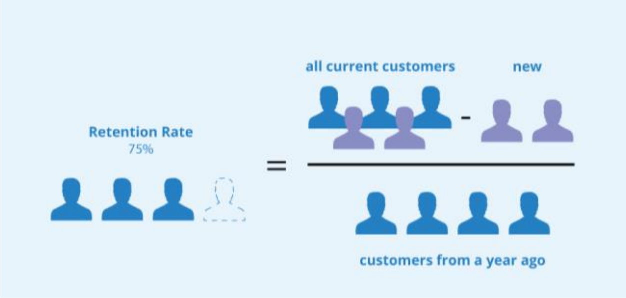
- Customer churn defined as “who left the service in the last month”

(target column “Churn”, if yes : 1, no: 0)

- 7043 instances (28% True, 72% False, Imbalanced)

- 19 features (16 categorical, 3 numerical)

“Who leaves and who remains?”

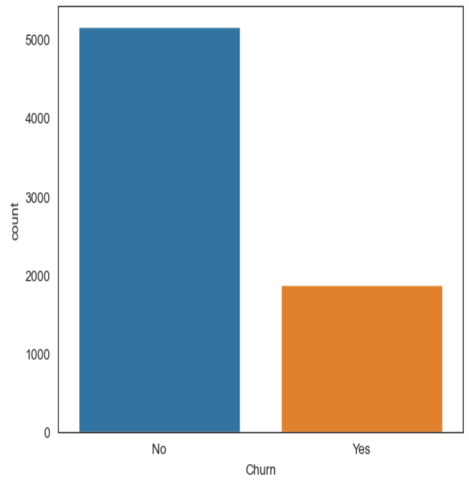


	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic
...	...	...	...	...	...	...	...	...	...
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes	Yes	DSL
7039	2234-XADUH	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic
7040	4801-JZAZL	Female	0	Yes	Yes	11	No	No phone service	DSL
7041	8361-LTMKD	Male	1	Yes	No	4	Yes	Yes	Fiber optic
7042	3186-AJIEK	Male	0	No	No	66	Yes	No	Fiber optic

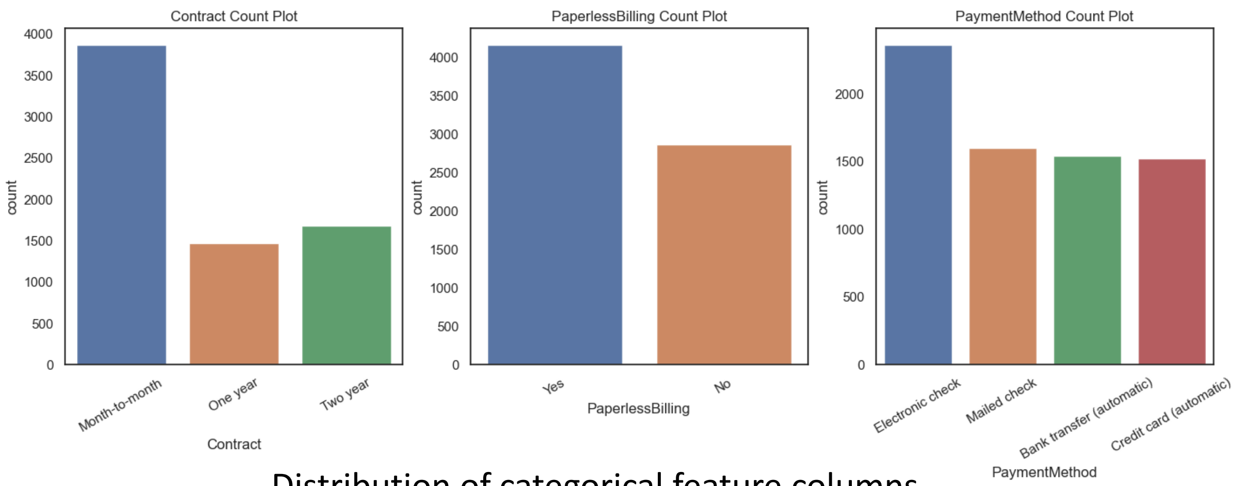
7043 rows × 21 columns

Data Exploration

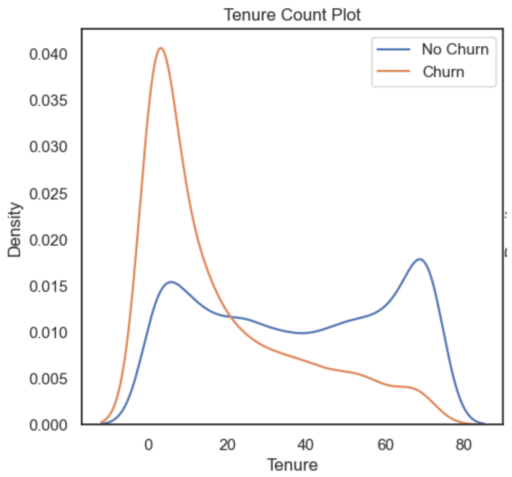
- Imbalanced data (28% True, 72% False)
- 19 Features – 3 types of info
  - Demographic info: Gender, SeniorCitizen, Partner, Dependents
  - Customer account info: Tenure, Contract, PaymentMethod, MonthlyCharges, TotalCharges
  - Other service info (signed up or not): PhoneService, MultipleLines, InternetService, StreamingTV etc.
- General Hypotheses
  - Those without partners, without dependents, senior are more likely to drop out
  - Those without security service, with Fiber optic Internet services are likely to drop out
  - Customers with month-to-month plans, using paperless billing are likely to drop out



Imbalanced data



Distribution of categorical feature columns

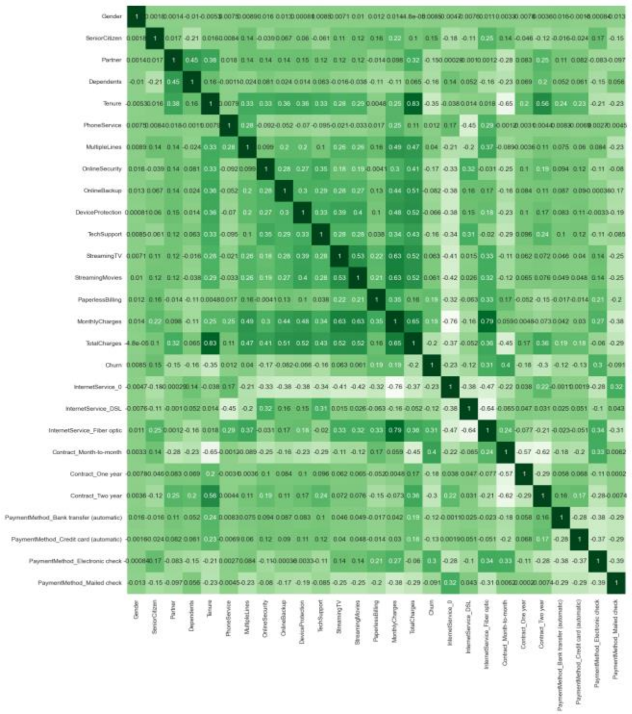


Distribution of numerical column

Data Pre-processing

- Relatively Clean data, pre-processing to streamline model training
- Pre-processing steps
  1. Merge categorical values without additional information
     - (Yes, No, No Internet service) → (Yes, No)
  2. Binary Encoding
     - Gender: Male → 0, Female →1
     - Binary categorical columns: Yes → 1, No → 0
  3. One-hot-encoding
     - Columns with more than 2 categorical values
  4. Scaler
     - Normalize numerical columns – MonthlyCharges, TotalCharges, Tenure

Correlation plot after pre-processing



	Gender	SeniorCitizen	Partner	Dependents	Tenure	PhoneService	MultipleLines
0	Female	0	Yes	No	1	No	No
1	Male	0	No	No	34	Yes	No
2	Male	0	No	No	2	Yes	No
3	Male	0	No	No	45	No	No
4	Female	0	No	No	2	Yes	No

Before pre-processing

	Gender	SeniorCitizen	Partner	Dependents	Tenure	PhoneService	MultipleLines
0	1	0	1	0	0.000000	0	0
1	0	0	0	0	0.464789	1	0
2	0	0	0	0	0.014085	1	0
3	0	0	0	0	0.619718	0	0
4	1	0	0	0	0.014085	1	0

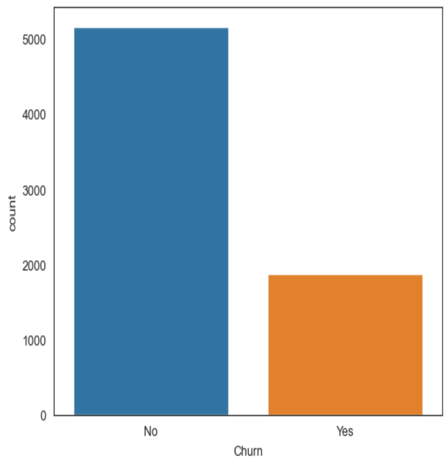
After pre-processing

Data Imbalance: Oversample vs Undersample

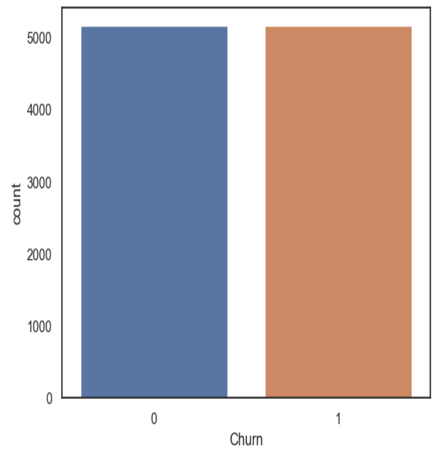
- Before splitting into train/test data, correct data imbalance
  - Small size of minority case (No-churn. Target value = 1) is problematic, as companies are more interested in the behavior of drop-out customers

	Oversampling	Undersampling
Chosen method	SMOTE	RandomUnderSampler
Cost to consider	Redundant data	Too small data size

- In general, undersampling is preferred over oversampling. But concerns about too small data size
- Change in sample size
  - Original: (1, 0) = (1869, 5163)
  - Oversampling: (1, 0) = (5163, 5163)
  - Undersampling: (1, 0) = (1869, 1869)
- Decide on which data set to use after testing on the models



Imbalanced data



Balanced data
After oversampling

Models – RF, LR, XGBoost

- 3 Model types: Random Forest, Logistic Regression, XGBoost
- Model building/deployment process

  : Simillar for the 3 models

  1. **GridSearchCV**
     - For both over/undersampled data
  2. Find the best parameters for each model
     - In all 3 models, higher score with <u>oversampled</u> data
  3. Present model performance
     - **Classification report**
     - **Confusion Matrix**
     - **AUC of ROC curve**
  4. After 1-3: Compare three model performance

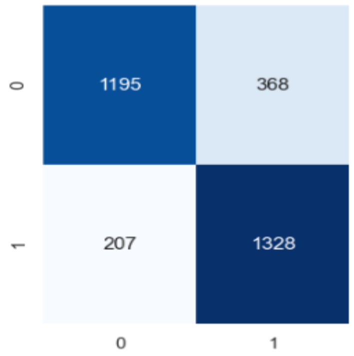
- Parameter grids for GridSearchCV
  1. Random Forest: n_estimators, max_depth, criterion
  2. Logistic Regression: penalty, C, max_iter, solver
  3. XGBoost: max_depth, learning_rate

[Random Forest Model]

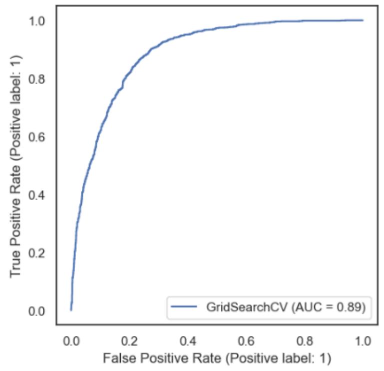
1. Classification report

	precision	recall	f1-score	support
0	0.85	0.76	0.81	1563
1	0.78	0.87	0.82	1535
accuracy			0.81	3098
macro avg	0.82	0.81	0.81	3098
weighted avg	0.82	0.81	0.81	3098

2. Confusion matrix



3. AUC – ROC curve



Models Performance

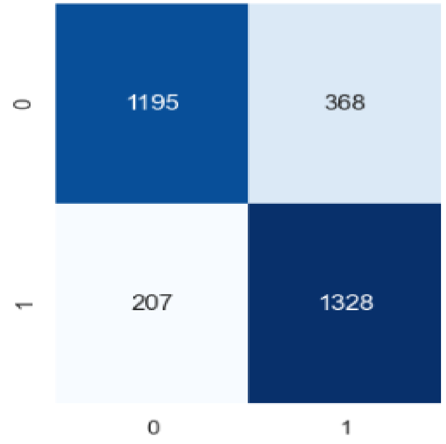
- Ranking of the model performance
  1. XGBoost
  2. Random Forest
  3. Logistic Regression

- Ranking was identical across three evaluation estimators

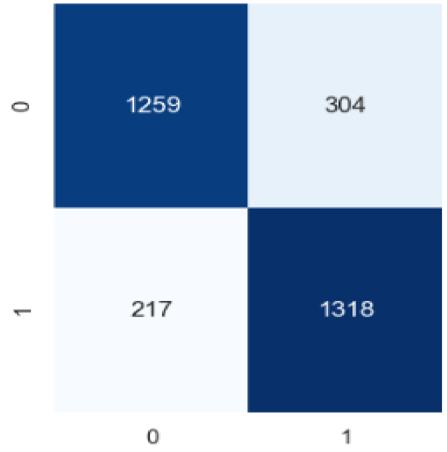
	Model	roc_auc_score	f1_score	accuracy_score
0	Random Forest	0.814851	0.822037	0.814396
1	Logistic Regression	0.811721	0.814838	0.811491
2	XGBoost	0.832067	0.834970	0.831827

[Detail]
RF: More TrueNegative vs. XGBoost: More TruePositive

Random Forest



XGBoost



Error Analysis

- Examine the incorrectly predicted data's distribution of each column with the overall data
- General Insights
  1. Models were not complicated enough to classify the case that does not follow the general trend
     - (e.g.,) There was a general trend that customers with dependents/partners are less likely to drop-out.
     - Noticeably unsuccessful in classifying customers without dependents/partners

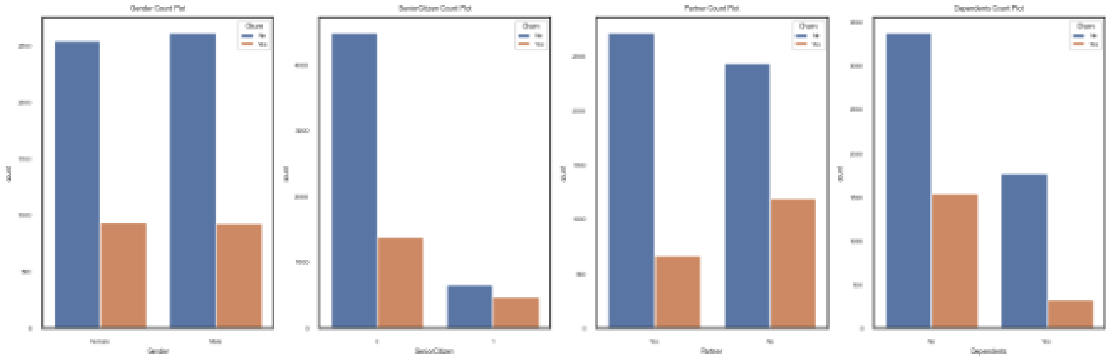
  2. Majority of columns were not significant enough to differentiate classes

     - When excluding the general trend detected in EDA, other columns does not seem to have explanatory power

  3. Three models are making common errors. Similar error rows

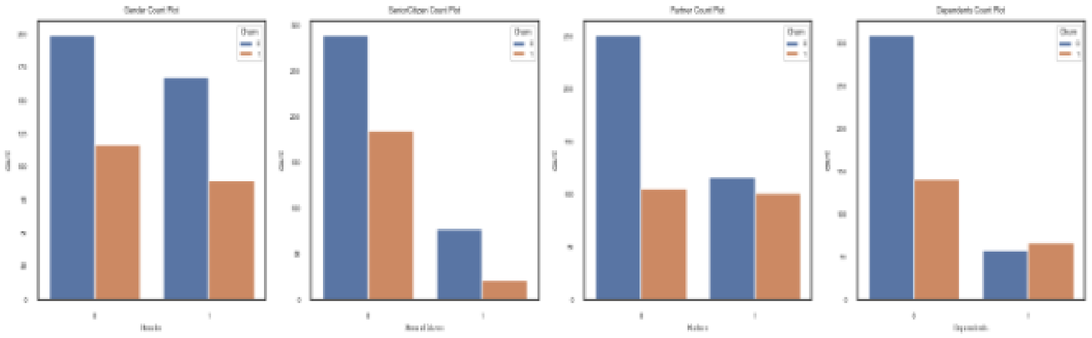
     - XGBoost was making less errors

Overall data – demographic columns



vs.

Incorrectly predicted data by RF– demographic columns



Further Task

Limitation of the project

1. Limited number of tested parameter grid
    - Including more parameters could have changed the result
2. Small data size for machine learning task
    - Oversampled data has 10326 instances
3. Comparison of three models
4. Technical aspect: Making model deployment and comparison into a function
    - Could have compare more models altogether