探索性分析

數據準備

Outlier

<https://scikit-learn.org/stable/modules/outlier_detection.html>

<https://www.kaggle.com/nareshbhat/outlier-the-silent-killer>

Imbalanced Data

Over-sampling (SMOTE)

SMOTE

Borderline-SMOTE

Borderline-SMOTE SVM

Adaptive Synthetic Sampling (ADASYN)

<https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>

Under-sampling

random

Tomek Links

Edited Nearest Neighbors

Neighborhood Cleaning

<https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/>

訓練、調整和評估機器學習模型

Cross Validation

沒有status的做semi-supervise

空值處理(邏輯補值方法)

中間值 平均 中位數

Feature篩選(PCA)

~~One hot model~~

分類模型：

XGBoost

<https://arxiv.org/abs/1603.02754>

<https://www.zhihu.com/question/58883125>

**random forests**

<https://web.archive.org/web/20080622230434/http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm>

**Support Vecor Machine (SVM)**

Decision Tree

GBM

Logistic regression

<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6>

<https://towardsdatascience.com/customer-churn-in-telecom-segment-5e49356f39e5>

demographics.csv: demographical information about customers

* CustomerID: A unique ID that identifies each customer.
* Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.
* Gender: The customer’s gender: Male(0), Female(1)
* Age: The customer’s current age, in years, at the time the fiscal quarter ended.
* Senior Citizen: Indicates if the customer is 65 or older: Yes, No
* Married: Indicates if the customer is married: Yes(1), No(0)
* Dependents: Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.
* Number of Dependents: Indicates the number of dependents that live with the customer.
* Age → Under 30, Senior Citizen(above 65)
* Number of Dependents → Dependents+

location.csv: contains geographical information about customers (only US data?)

* CustomerID: A unique ID that identifies each customer.
* Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.
* Country: The country of the customer’s primary residence.
* State: The state of the customer’s primary residence.
* City: The city of the customer’s primary residence.
* Zip Code: The zip code of the customer’s primary residence.
* Lat Long: The combined latitude and longitude of the customer’s primary residence. 經緯度組合
* Latitude: The latitude of the customer’s primary residence.
* Longitude: The longitude of the customer’s primary residence.
* (查美國各州的zipcode對應?)

population.csv: contains population information of each area

* ID: A unique ID that identifies each row.
* Zip Code: The zip code of the customer’s primary residence.
* Population: A current population estimate for the entire Zip Code area.

satisfaction.csv: contain satisfaction score from survey

* CustomerID: A unique ID that identifies each customer.
* Satisfaction Score: A customer’s overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied). (NaN implement to 3?)

services.csv: contains information about the services that a customer used

* CustomerID: A unique ID that identifies each customer.
* Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.
* Quarter: The fiscal quarter that the data has been derived from (e.g. Q3).
* Referred a Friend: Indicates if the customer has ever referred a friend or family member to this company: Yes, No
* Number of Referrals: Indicates the number of referrals to date that the customer has made.
* Tenure in Months: Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above. 截至本季末, 客戶續約月數
* Offer: Identifies the last marketing offer that the customer accepted, if applicable. Values include None, Offer A, Offer B, Offer C, Offer D, and Offer E. 簽約種類
* Phone Service: Indicates if the customer subscribes to home phone service with the company: Yes, No
* Avg Monthly Long Distance Charges: Indicates the customer’s average long distance charges, calculated to the end of the quarter specified above.
* Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No
* Internet Service: Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.
* Avg Monthly GB Download: Indicates the customer’s average download volume in gigabytes, calculated to the end of the quarter specified above.
* Online Security: Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No
* Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No
* Device Protection Plan: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No
* Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No
* Streaming TV: Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.
* Streaming Movies: Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.
* Streaming Music: Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.
* Unlimited Data: Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No
* Contract: Indicates the customer’s current contract type: Month-to-Month, One Year, Two Year.
* Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No
* Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
* Monthly Charge: Indicates the customer’s current total monthly charge for all their services from the company.
* Total Charges: Indicates the customer’s total charges, calculated to the end of the quarter specified above.
* Total Refunds: Indicates the customer’s total refunds, calculated to the end of the quarter specified above.
* Total Extra Data Charges: Indicates the customer’s total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above.
* Total Long Distance Charges: Indicates the customer’s total charges for long distance above those specified in their plan, by the end of the quarter specified above.

status.csv: contains imformation about customers' status

* CustomerID: A unique ID that identifies each customer.
* Churn Category: customer’s reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price, No Churn. When they leave the company, some customers are asked about their reasons and classified them into 5 category. No Churn indicates the customer still stays in the company.

Test\_IDs.csv: contains Customer ID in the testing dataset

Train\_IDs.csv: contains Customer ID in the training dataset