**Telco Churn Prediction**

1. **Main Objective**

Attemp to predict customer whether the customer will leave the company (churned) or not using classification techniques with the focus on **prediction task**.

In this report, the main goal is breakdown analysis of classification task of telco company’s customer churn problem.

1. **Data Definition and Description**

This dataset was provided by IBM course material, containing customer behaviors and factor variables that might cause churn. The customer ids might not actual customer id and the dataset information is intended for educational purposes only.

The dataset contain 7043 observations and 21 attributes (variables), 16 variables after feature selection.

**Data description**

'months' : total customer duration on staying with the company

'offer' : customers segmentation with several product offering

'multiple': Whether the customer has multiple lines or not

'internet\_type': customer internet type

'gb\_mon': customer speding on internet usage(Gb) per month

'security': whether customer use security online service or not

'backup': whether customer use backup service or not

'protection': device protection

'support': premium tech support plan

'unlimited': unlimited internet plan

'contract': contract type of customer

'paperless': customer use paperless billing

'payment': how customer pay billing

'monthly': monthly charge

'satisfaction': customer satisfaction level

'churn\_value': customer churn status

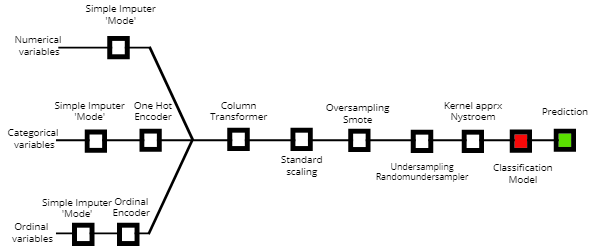
**Research Objective**

Build a model that yield best accuracy without suffering accuracy paradox and great performace in terms of training and predicting speed

1. **Pre-processing and Feature Engineering Plan**

* Inspect the columns and its data type
* Inspect null value
* Inspect duplicate
* Fix the column data type and names
* Create custom function to get column statistic for detecting and outlier treatment
* Inspect outliers and treat the outliers with reasonable value
* Due to prediction main objective, variable transformation will be performed
* Standardize the numerical feature
* One hot encode to the categorical feature
* Ordinal encoder to the ordinal feature
* Create machine learning pipeline for further transformation;

**Pipeline Diagram**

****

1. **Model Selection and Training Results**

To achieve best prediction accuracy, we will conduct three different Models which are Support Vector Machine, Logistic Regression, and Adaboostclassifier. The model training conducted with ;

* Pipeline technique
* cross-validation with 5 k-fold
* The model training use train-test split with 700 sample for test set
* Train test Random state: 42
* Precision, recall, f1 score, accuracy, and area under the curve
* Grid search for searching best parameter

**Modeling**

1. **Support Vector Machine**

**To search parameter** : {kernel = polynomial, rbf; degree = 2,4,6,8 ; C = 0.01,0.5,1,3,5,10}

**Best parameter used** ;

* Kernel : radial basis function
* Alpha (regularization) : 1.0
* Gamma: ‘scale’
* Combined with Nystroem kernel approximation = rbf kernel
* Model explainability : Very Low (Black box model)

1. **Logistic Regression**

**To search parameters** {penalty = none,elasticnet (L3) ; C= 0.01,0.5,1,3,5,10 ; l1\_ratio= 0,0.25,0.5,0.75,1}

**Best parameter used** ;

* Best penalty : none (no penalty)
* With Nystroem Kernel approximation = rbf kernerl

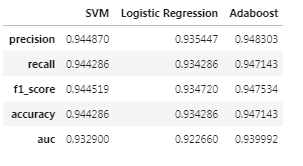
1. Adaboost classifier

Summary performance of Adaboost on hold-out test set ;

**To search parameters** {n\_estimator : 50,100,150,200,300}

* Base model = Decision Tree classifier
* Best number of estimator : 150
* Learning rate = 1
* Random state = 42
* Without nystroem

1. **Model Selection, Conclusion, and Summary**



Based on the performance metrics it is highly **recommended to Adaboost Classifier** with the best results below:

* Precision 94.8%
* Recall 94.7%
* Harmonic score f1 94.7%
* Accuracy 94.7%
* Auc 94.7%

From now on, this model will be mentioned as choosen model.

Due to prediction accuracy main objective, the model training will focus on how to create a model with high prediction accuracy, without suffering accuracy paradox as well. The **model has low interpretability** due to combined oversampling and undersampling techniques that very important to drive the model accuracy and no longer the model will be **a black box model**.

Based on the parameters used and other model comparison, we can conclude that below conlusion will drive model accuracy;

1. Class rebalancing techniques such as SMOTE and Random Under Sampling
2. Optimized number of decision tree weak learners
3. Using kernel approximation might increase model training performance in this case using nystroem will drop a little bit of class f1 score
4. While our model has great performance and accuracy, it will be good at predicting the customer whethe it will churn or not. The prediction results may be useful to the company to prevent the potential churned customer with possible action such as:

* Direct offer/promotional
* Personalized marketing
* Budget recalculating
* Company’s revenue prediction

1. **Further Development**

With dataset open-sourcity and broader techniques, any researcher can expand this report or conduct different research such as re-conducting research for interpretability purpose, exploratory data analysis on this dataset, and many more. As the author, i would suggest we should conduct what is the biggest reason behind churned customers

**References**

Author’s notebook and source code

<https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/f3d9887f-3a1b-43f1-9832-6600973ec610/view?access_token=3657fea0a0d5810f91dfd3c280c29ace50b4ec2babb002a82b0128f1edff2408>

similiar dataset

<https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>