**Telco Churn Prediction**

**Technical Report**

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.

**Research Objective**

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

1. **Data Definition and Description**

This dataset was provided by IBM data exchange, containing customer service behaviors, demographics 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 54 attributes (variables), 24 variables after feature selection. Missing value on churn category 5174 samples, churn reason 5174,

No duplicate sample, and raw format as pkl (pickle object).

**Action taken for feture selection**

With Churn label column as dependent variable and has two class categorical, feature selection applied to raw rata to improve model performace and choose only important features. business understanding feature selection and hypothesises testing tools method is used for this research.

**Used Data Description**

Under 30 : whether the customer is aged under 30 or not, Yes or No

Married : Indicates if the customer is married: Yes, No

Number of Dependents : Indicates the number of dependents that live with the customer.

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.

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.

Multiple Lines : Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No

Internet Service : wheter customer using internet service or not, Yes or No

Internet Type : Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

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

Total Charges : Indicates the customer’s total charges, calculated to the end of the quarter specified above.

Total Revenue : indicates the total revenue made from customers

Satisfaction Score : A customer’s overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied).

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

Feature selection pre-processing, for the long run purposes we create new temporary column encoded\_label from churn label. The new column is the result of binary encoded churn label containing 0 (No) and 1 (Yes).

**Correlation coefficient test**:

* Pearson correlation coefficient

Pearson coefficient interpretation baseguide: Statistic without math for psychology by Dancey and Reidy

* Point biserial correlation (pearson correlation)

**Hypothesis testing tools**:

* Chisquare contingency

H0 : the variable is independent to target (no significant correlation)

H1 : the variable is correlated to target

Alpha : 5%

Correlation coefficient and hypothesis testing against churn\_label/encoded churn results:

* age, under 30, and senior variables are referring to the same information. to avoid multicollinearity we will use only under 30. Under 30 tested with Chisquare contingency resulting P-value: 0.00032, reject H0 and keep the feature for modeling.
* Married : tested with chisquare contingency resulting p-value 1.95x10-33, reject H0 and keep the feature for modeling
* Number of dependents : tested with pearson correlation resulting coefficient: -0.218, resulting coef is in acceptable value and we keep the feature for modeling
* Tenure in months : tested with pearson correlation resulting coefficient: -0.315, resulting coef is in acceptable value and we keep the feature for modeling
* Number of referrals : tested with pearson correlation resulting coefficient: -0.286, resulting coef is in acceptable value and we keep the feature for modeling
* Offer : tested with chisquare contingency resulting p-value 1.71x10-95, reject H0 and keep the feature for modeling
* Phone service : tested with chisquare contingency resulting 0.90, fail to reject H0 and feature not used for modeling
* Avg Monthly Long Distance Charges : tested with pearson correlation resulting coefficient: 0.0081, resulting coef is not in acceptable value and feature not used for modeling
* Multiple lines: tested with chisquare contingency resulting p-value 0.023, reject H0 and keep the feature for modeling
* Internet service : tested with chisquare contingency resulting p-value 6.89x10-78, reject H0 and keep the feature for modeling
* Internet type : tested with chisquare contingency resulting p-value 6.18x10-136, reject H0 and keep the feature for modeling
* Avg Monthly GB Download tested with pearson correlation resulting coefficient: 0.048, resulting coef is not in acceptable value and feature not used for modeling
* Online Security : tested with chisquare contingency resulting p-value 1.51x10-43, reject H0 and keep the feature for modeling
* Online backup : tested with chisquare contingency resulting p-value 1.11x10-9, reject H0 and keep the feature for modeling
* Device protection plan : tested with chisquare contingency resulting p-value 3.31x10-6, reject H0 and keep the feature for modeling
* Premium tech support : tested with chisquare contingency resulting p-value 3.25x10-40, reject H0 and keep the feature for modeling
* Streaming Tv : tested with chisquare contingency resulting p-value 1.15x10-05, reject H0 and keep the feature for modeling
* Streaming movies : tested with chisquare contingency resulting p-value 2.46x10-5, reject H0 and keep the feature for modeling
* Streaming music tested with chisquare contingency resulting p-value 0.005, reject H0 and keep the feature for modeling
* Unlimited data tested with chisquare contingency resulting p-value 3.74x10-41, reject H0 and keep the feature for modeling
* Contract tested with chisquare contingency resulting p-value 3.77x10-309, reject H0 and keep the feature for modeling
* Paperless billing tested with chisquare contingency resulting p-value 6.91x10-55, reject H0 and keep the feature for modeling
* Payment method tested with chisquare contingency resulting p-value 6.31x10-70, reject H0 and keep the feature for modeling
* Monthly charge and total charges refering to the same information, only total charges used to avoid multicollinearity. tested with pearson correlation resulting coefficient: -0.198, resulting coef is in acceptable value and we keep the feature for modeling
* Total refunds tested with pearson correlation resulting coefficient: -0.03, resulting coef is not in acceptable value and not keeped the feature for modeling
* Total revenue tested with pearson correlation resulting coefficient: -0.22, resulting coef is in acceptable value and we keep the feature for modeling
* Satisfaction score tested with pearson correlation resulting coefficient: -0.75, resulting coef is in acceptable value and we keep the feature for modeling
* Customer’s location information does not relate to solve problem objective
* **Data type and the variable type inspection**

Under 30 : categorical

Married : categorical

Number of dependents : numerical

Number of referral : numerical

Tenure in months : numerical

Offer : categorical

Multiple lines : categorical

Internet service : categorical

Internet type : categorical

Online security: categorical

Online backup : categorical

Device protection plan : categorical

Premium tech support : categorical

Streaming tv : categorical

Streaming movies : categorical

Streaming music : categorical

Unlimited data : categorical

Contract : ordinal categorical

Paperless billing : categorical

Payment method : categorical

Total charges : numerical

Total revenue: numerical

Satisfaction score : ordinal

Churn label : target, categorical

* **Inspect null value**

No null value detected from selected/filtered data

* **Inspect duplicate**
* No duplicate value detected from selected/filtered data
* **Fix the column data type and names**

No improper format and names found

* **Inspect outlier and outlier treatment**

Outlier detection tool : Grubbs outlier hypothesis testing

H0 there is no outliers

H1 there is at least 1 outlier

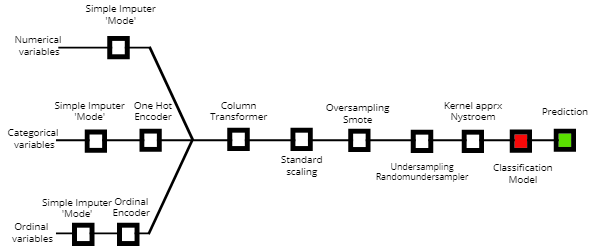
Alpha 5%

Result: number of dependents has 3 outliers

action taken : leave it to be

* **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 model parameter on Support Vector Machine. The model training conducted with ;

* Pipeline technique
* cross-validation with 5 k-fold
* The model training use train-test split with 10% sample for holdout test set
* Train test Random state: 42
* Precision, recall, f1 score, accuracy, and area under the curve metrics
* Grid search for searching best parameter

**Modeling**

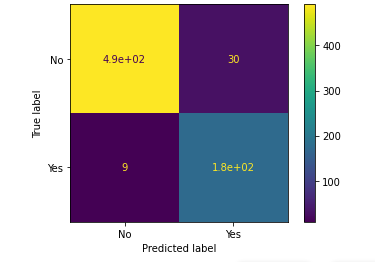
1. **Support Vector Machine**

**To search parameter** : {kernel = polynomial, rbf; degree = 3,6,8 ; C = 0.01,10,50,250,500,1000 , gamma= ‘auto’, 0.01, 10}

**Best parameter used** ;

* Kernel : radial basis function
* Alpha (regularization) : 100
* Gamma: 0.01
* Model explainability : Very Low (Black box model)
* Degree: 3
* Class\_weight : balanced

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



Confusion matrix of holdout test set.

True label : 670

False label : 39

Based on the performance metrics it is highly **recommended to Support vector machine** with the best results below:

* Weighted Precision 95%
* Weighted Recall 94%
* Harmonic score f1 93%
* Accuracy 94.7%
* Auc 99%

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** and 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 will drive model precision and recall score
2. 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://github.com/zylbergs/Project_Portofolio/blob/main/TelcoChurn_classification/telcochurn_prediction_from_watson.ipynb>

similiar dataset

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