# Telco churn prediction model

**Peerapat.t, Data scientist**
For project's material please visit me on : github.com/peerapat-t

# Table of content

# 1. Business problem

### 1.1 Cost of acquiring a new customer

- The cost of retaining a customer is typically much lower than the cost of acquiring a new customer.
- For example, a study by Bain & Company found that it costs 5 to 7 times more to acquire a new customer than it does to retain an existing customer.
- This is because existing customers are already familiar with your product or service and they are more likely to continue doing business with you.

### 1.2 Why does churn prediction matter?

- Churn prediction can help businesses identify customers who are at risk of leaving.
- This allows businesses to take action to retain those customers, such as offering them discounts or special promotions.
- Churn prediction can also help businesses to identify trends in customer behavior that may lead to churn, so that they can address those issues before they cause customers to leave.

# 2. How to solve this problem

**2.1 Do nothing**

- Business not take any action to retain customers.

**2.2 Retain all customers**

- Business take action to retain every customer.

**2.3 Use a churn prediction model**

- Business use churn prediction to predicts who are at risk of churning.

| Option | Cost | Effectiveness |
|---|---|---|
| Do nothing | Least expensive | Least effective |
| Retain all customers | Most expensive | Most effective |
| Use a churn prediction model | More cost-effective than retaining all customers | More effective than doing nothing |

# 3. Data

### 3.1 Target

- There are 7,032 customers in the dataset, including churn and not-churn customers.
- Churn customers are those who have unsubscribed within the last month.

### 3.2 Features

- The dataset contains 22 features, which can be categorized into four groups:
    - Services that each customer has signed up for: This includes phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information: This includes how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
- Demographic info about customers: This includes gender, age range, and if they have partners and dependents.

# 4. Solution

### 4.1 Churn prediction model performance

| Option | Total (Test) | Actual churn | Overspend (FP) | Save (TP) | Gains (or Loss) |
|---|---|---|---|---|---|
| Do nothing | 1,407 | 374 | 0 | 0 | -187,000 |
| Retain all | 1,407 | 374 | 1,033 | 374 | 83,700 |
| Churn prediction model | 1,407 | 374 | 307 | 295 | 110,700 |

- Utilizing data from the test set involving approximately 1,407 customers, the churn prediction model yielded a gain of 3.38 times when compared to the 'do nothing' program, and 0.32 times when compared to the 'retain all' program.
- These calculations are based on the assumption that the cost of promotion is 100, the cost of loss is 500, and the savings per customer is 500."

# 4. Solution

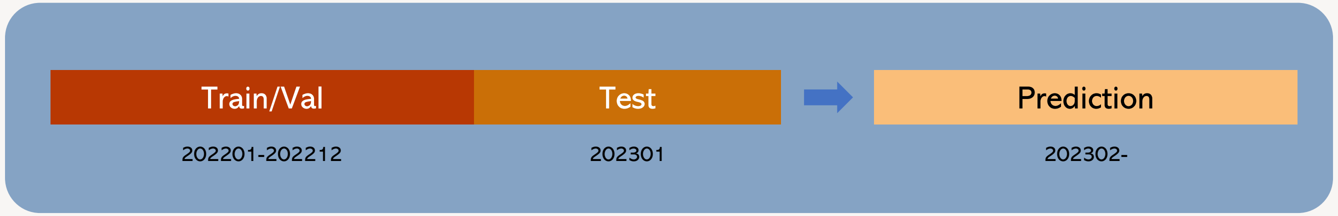
### 4.2 Promotion/campaign to retain customers

- **For customers who have signal of churning and using a top-up package,** we should introduce promotions to transition them to a contract package. Additionally, provide a limited-time offer like free 10GB internet for 6 months.
- **For customers who have signal of churning and they are elder,** we should offer a budget-friendly package accompanied by a message like "Stay connected with your number.
- **For customers who have signal of churning and using a low-price package,** we should be presented with a comparison of our package against competitors' offerings, followed by a strategic response.
- **For customers who have signal of churning and no current internet service,** we should introduce an affordable internet package.
- **For customers who have signal of churning and still pay bills physically,** we should encourage them to explore more convenient payment options, such as automatic credit card billing.

# 4. Solution

### 4.3 Use of model in the future

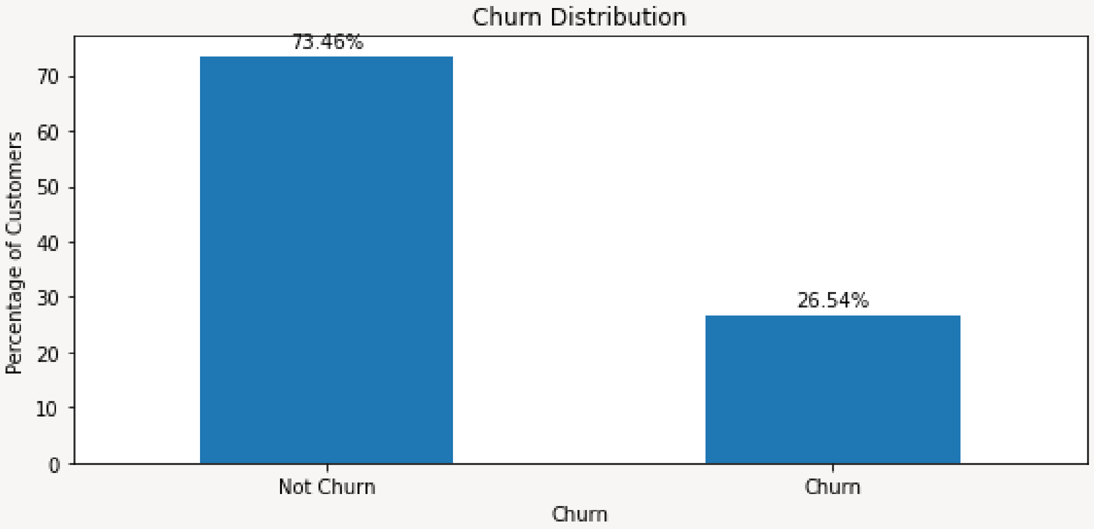
- Predict customer who are at risk to churn in the next month using the model and send leads to marketing team for retention campaign.



# End of business session

# 4. Model

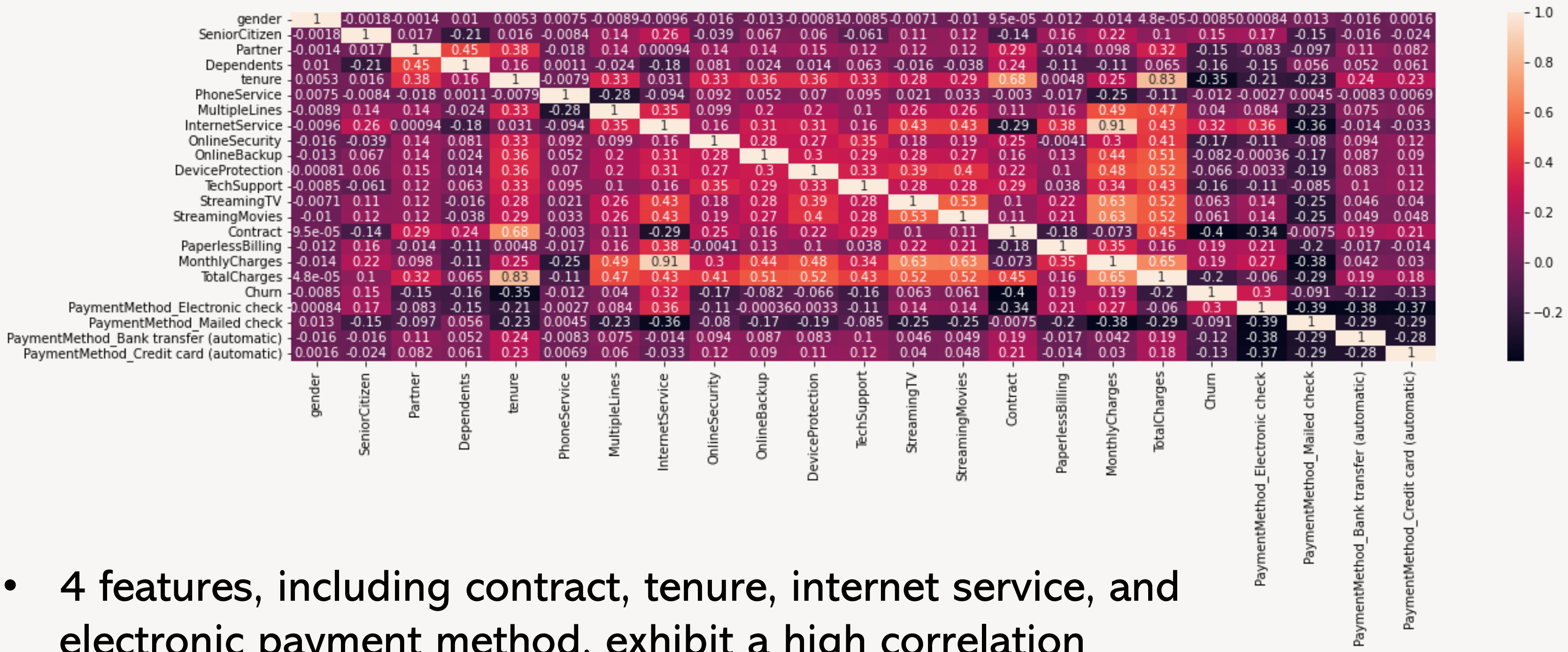
## 4.1 Data understanding




- If we look at the target distribution, There are only 27% of churn customers that the data is imbalanced.
- Resampling method or threshold tuning should be applied to handle this problem.
- Metrics like accuracy can be misleading and ineffective. Instead, it is essential to explore alternative evaluation measures, such as precision, recall, F1-score, or AUC-ROC.

# 4. Model

## 4.1 Data understanding (cont.)




| | |
|---|---|
| Churn | 1.000000 |
| Contract | 0.396150 |
| tenure | 0.354049 |
| InternetService | 0.316350 |
| PaymentMethod_Electronic check | 0.301455 |
| TotalCharges | 0.199484 |
| MonthlyCharges | 0.192858 |
| PaperlessBilling | 0.191454 |
| OnlineSecurity | 0.171270 |
| TechSupport | 0.164716 |
| Dependents | 0.163128 |
| SeniorCitizen | 0.150541 |
| Partner | 0.149982 |
| PaymentMethod_Credit card (automatic) | 0.134687 |
| PaymentMethod_Bank transfer (automatic) | 0.118136 |
| PaymentMethod_Mailed check | 0.090773 |
| OnlineBackup | 0.082307 |
| DeviceProtection | 0.066193 |
| StreamingTV | 0.063254 |
| StreamingMovies | 0.060860 |
| MultipleLines | 0.040033 |
| PhoneService | 0.011691 |
| gender | 0.008545 |

- 4 features, including contract, tenure, internet service, and electronic payment method, exhibit a high correlation (greater than 0.3 in absolute terms) with the target variable."

# 4. Model

## 4.2 Technique used

| Step | Topic | Cost |
|---|---|---|
| 1 | Dataset | • Split the data into train, test, and validation sets.<br>• Use the train set to train the model, the validation set to tune the threshold, and the test set to evaluate the model. |
| 2 | Normalized | • Use Min-Max scaling to normalize the data. |
| 3 | Resampling | • Try oversampling (SMOTE) and undersampling (random undersampling). |
| 4 | Model | • Try random forest, LightGBM, and XGBoost. |
| 5 | Hyperparameter tuning | • Use RandomizedCV to find the best hyperparameters for each model. |
| 6 | Threshold tuning | • Use cost-sensitive learning to tune the threshold.<br>• Assign a cost of 5 times more to acquiring new customers than to retaining existing customers. |
| 7 | Interpreting | • Use SHAP values. |

# 4. Model

## 4.3 Model evaluation

| | Model | Observation | TP | TN | FP | FN | Precision | Recall | AUCROC |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Dummy Model (All Churn) | 1407 | 374 | 0 | 1033 | 0 | 0.265814 | 1.000000 | 0.500000 |
| 1 | Dummy Model (All Not Churn) | 1407 | 0 | 1033 | 0 | 374 | NaN | 0.000000 | 0.500000 |
| 2 | Random Forest (SMOTE) | 1407 | 252 | 812 | 221 | 122 | 0.532770 | 0.673797 | 0.824113 |
| 3 | LightGBM (SMOTE) | 1407 | 256 | 808 | 225 | 118 | 0.532225 | 0.684492 | 0.819451 |
| 4 | XGBoost (SMOTE) | 1407 | 289 | 742 | 291 | 85 | 0.498276 | 0.772727 | 0.828354 |
| 5 | Random Forest (RUS) | 1407 | 294 | 729 | 304 | 80 | 0.491639 | 0.786096 | 0.828529 |
| 6 | LightGBM (RUS) | 1407 | 295 | 726 | 307 | 79 | 0.490033 | 0.788770 | 0.828037 |
| 7 | XGBoost (RUS) | 1407 | 306 | 688 | 345 | 68 | 0.470046 | 0.818182 | 0.826978 |

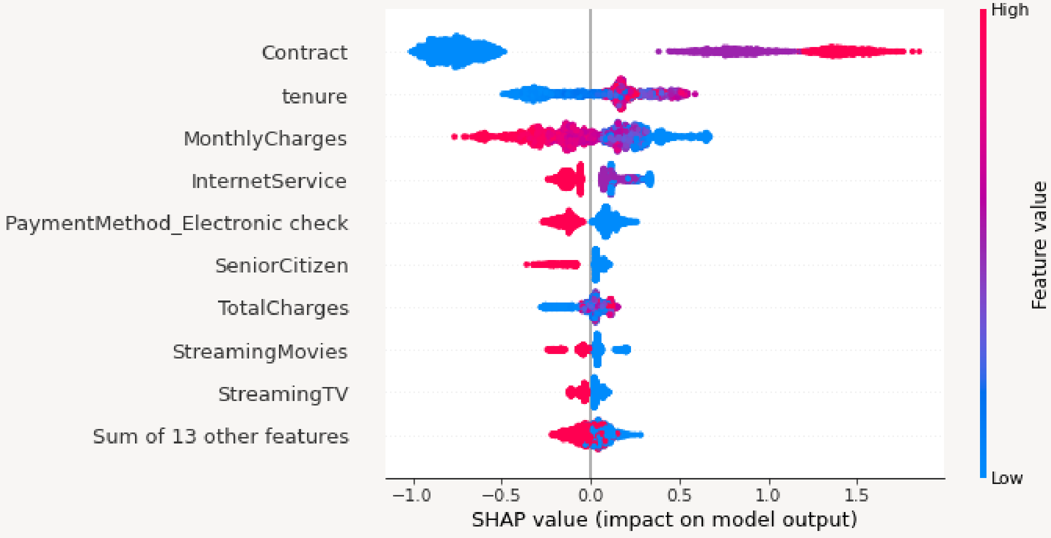
| | Model | Best Threshold | Validation Gain | Test Gain |
|---|---|---|---|---|
| 0 | Dummy Model (All Churn) | 0.00 | 69300 | 83700 |
| 1 | Dummy Model (All Not Churn) | 0.00 | -151500 | -187000 |
| 2 | Random Forest (SMOTE) | 0.18 | 95500 | 108400 |
| 3 | LightGBM (SMOTE) | 0.25 | 95900 | 102400 |
| 4 | XGBoost (SMOTE) | 0.19 | 97900 | 108000 |
| 5 | Random Forest (RUS) | 0.28 | 96500 | 109400 |
| 6 | LightGBM (RUS) | 0.25 | 94700 | 110700 |
| 7 | XGBoost (RUS) | 0.31 | 96800 | 108200 |

- Before performing threshold tuning, XGBoost (RUS) seems good because it has a high AUCROC and the highest Recall.
- However, prior to that, LightGBM (RUS) is the better model as it achieves maximum gains.
- Ultimately, for the final decision, we choose LightGBM (RUS) to deploy in our system."

# 4. Model

## 4.4 Feature importance (cont.)

| Features | Sign | Meaning |
|---|---|---|
| Contract | + | Customers on contract or top-up plans are more likely to churn, possibly due to the ease of changing numbers for top-up customers. |
| Tenure | + | Long-time customers have a higher churn probability, potentially due to elderly individuals reducing phone usage to cut expenses. |
| Monthly charges | - | Customers paying lower charges are more likely to churn, influenced by industry trends towards appealing, low-priced packages from other operators. |
| Internet service | - | Customers without internet in their package are at a higher churn risk, as competitors offer inclusive internet services, reflecting its growing importance. |
| Payment method (electronic check) | - | Non-users of electronic payment for bills are more likely to churn, possibly due to the inconvenience of payment, prompting a switch to other providers. |



# 5. Future work

- Improve model performance by creating more features and performing feature engineering.
- Experiment with different machine learning models, such as SVC and deep learning.
- Tune hyperparameters using sequential search techniques, such as Optuna.
- Segment customers using clustering and retain them with personalized promotions.

# End of this presentation

# 6. About me

- Peerapat Tancharoen holds a Bachelor's degree in Economics from Srinakharinwirot University, graduating with first honors and a GPA of 3.67. He also earned a Master's degree in Economics from Thammasat University, achieving a GPA of 3.98.
- With four years of experience as a Data Scientist at Kasikorn Asset Management, Peerapat is dedicated to utilizing machine learning models to address intricate business challenges, notably in the realm of marketing. His passion lies in leveraging data-driven solutions to drive impactful outcomes.


**Contact me on**

- linkedin.com/in/peerapat-tancharoen-664759220/
- peerapat.tcr@gmail.com