**PRD: Application to Predict** Customer Churn in Telco Industies

# Problem Alignment

## The Problem

*Describe the problem (or opportunity) you’re trying to solve. Why is it important to our users and business? What insights are you operating on? And if relevant, what problem are you not intending to solve?*

Retaining customers is one of the most critical challenges in the maturing mobile telecommunications service industry. Customer retention is one of the primary KPI for companies with a subscription-based business model. Competition is tough particularly in telco market where customers are free to choose from plenty of providers. One bad experience and customer may just move to the competitor resulting in customer churn.

Predicting customer churn is a challenging but extremely important business problem especially in industries where the cost of customer acquisition (CAC) is high such as technology, telecom, finance, etc. The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for companies.

The primary objective of the customer churn predictive model is to retain customers at the highest risk of churn by proactively engaging with them. For example: Offer a gift voucher or any promotional pricing and lock them in for an additional year or two to extend their lifetime value to the company.

## High-Level Approach

*Describe briefly the approach you’re taking to solve this problem. This should be enough for the reader to imagine possible solution directions and get a very rough sense of the scope of this project.*

Customer churn is the loss of clients or customers. In order to avoid losing customers, a company needs to examine why its customers have left in the past and which features are more important to determine who will churn in the future. Our task is therefore to predict whether customers are about to churn and which are the most important features to get that prediction right. As in most prediction problems, we will use machine learning.

The data set used in this project is publicly available dataset provided by kaggle telco customer churn that available in <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>. This dataset contains nineteen columns (independent variables) that indicate the characteristics of the clients of a fictional telecommunications corporation. The Churn column (response variable) indicates whether the customer departed within the last month or not. The class No includes the clients that did not leave the company last month, while the class Yes contains the clients that decided to terminate their relations with the company. The objective of the analysis is to obtain the relation between the customer’s characteristics and the churn.

In this project, we intend to create a churn prediction using *classification methods*. This “curn” is the target variable, which is numerical and contains values 2 classes. Values ‘Yes' represents our customer are churn (positive class) and ‘No’ value represents our customer are not churn (negative class).

Here, we compare various AI algorithms, including classical Machine Learning, deep learning architecture and an ensemble of algorithms on the publicly available dataset provided by kaggle telco customer churn cases dataset that available in <https://www.kaggle.com/datasets/blastchar/telco-customer-churn> . Our ML model will be published as a backend API Machine Learning. For the last part, our ML API should have a GUI frontend so that it will make it easier for the sales team to predict whether customers will be churn or not based on their demographic data.

## Goals & Success

*What does success look like? What metrics are you intending to move? Explain why these metrics are important if not obvious.*

\*Convert business metrics to ML metrics. (exp: people likeliness to choose one restaurant →MAP or recall?

Insights on data and metrics such as confusion matrices, F1-score, precision, recall, receiver operating characteristic (ROC) curves will be demonstrated in our output metrics. We also consider converting probability measured results to suggest a decision to the customers whether to take the they will churn or not.

Evaluating the quality of the model is a fundamental part of the machine learning process. The most used performance evaluation metrics are calculated based on the elements of the confusion matrix.

* Accuracy: It represents the proportion of predictions that were correctly classified. Accuracy is the most commonly used evaluation metric; however, it is important to bear in mind that accuracy can be misleading when working with imbalanced datasets like in our case.



* Sensitivity: It represents the proportion of positive samples (churn) that are identified as such.



* Specificity: It represents the proportion of negative samples (no churn) that are identified as such.



* Precision: It represents the proportion of positive predictions that are actually correct.



We can calculate the evaluation metrics manually using the numbers of the confusion matrix. Alternatively, Scikit-learn has already implemented the function classification\_report that provides a summary of the key evaluation metrics. The classification report contains the precision, sensitivity, f1-score, and support (number of samples) achieved for each class.

# Solution Alignment

## Key Solution

*Give an overview of what the dataset is. Provide an organized list of features that may contribute to the modeling. Discuss the constraint*

Since the target variable only contains 2 classes, this case is considered as a binary classification problem. We used logistic regression (classical Machine Learning) as a baseline classifier and interpretable models. Furthermore, various AI algorithms, both deep learning architecture, and an ensemble of algorithms will be applied to predict “churn” in telco churn dataset. Imbalance handling for the dataset is also considered in our machine learning processing using SMOTE and Undersampling Methods.

## Key Flows

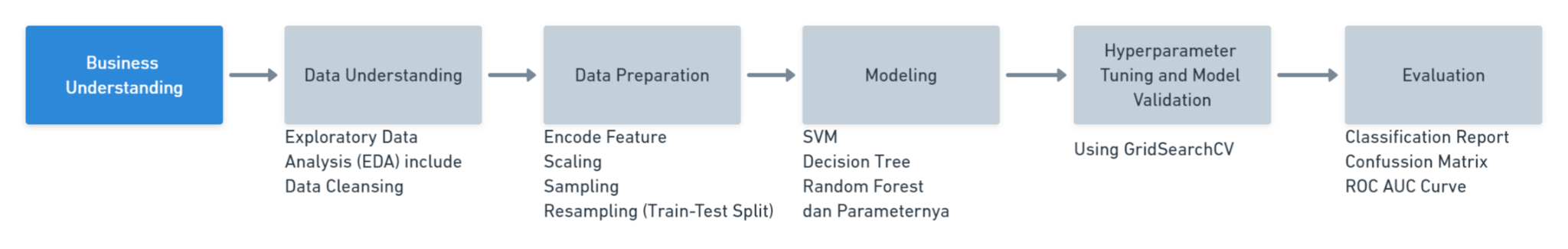
*Show some mocks/embeds of the experience. Link to any other documentation as necessary.*

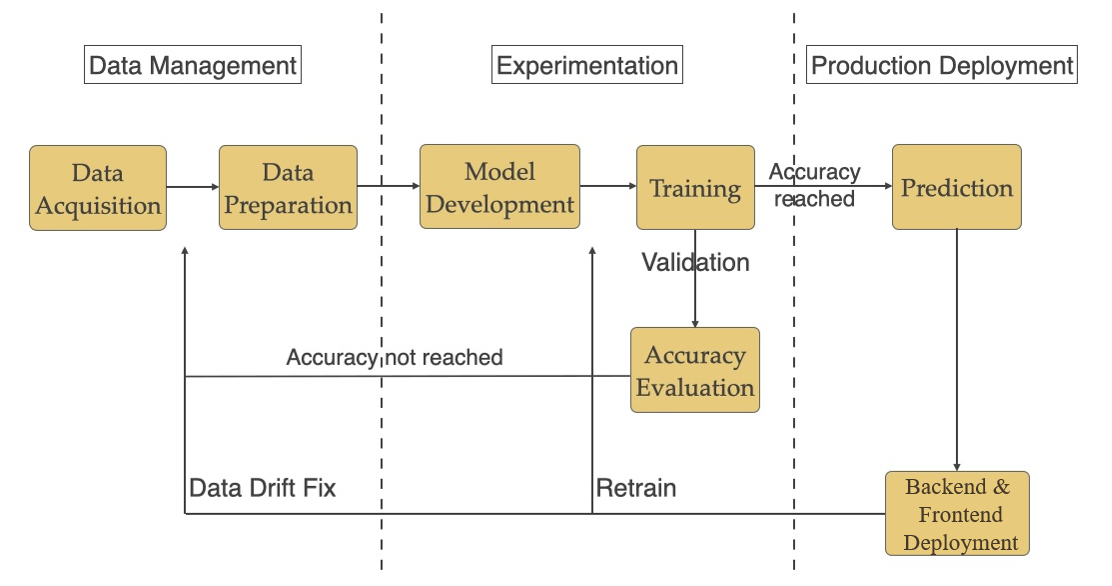
Tools : API Builder (Flask) & ML Library in Python

Data Prep : Imbalance data handling, remove duplicates and Missing value handling

Model : Logistic regression (classical Machine Learning) as a baseline classifier and interpretable models. Ensemble of algorithms : RF, XGBoost, Catboost. Deep learning architecture Modeling

Deployment Environments : API ML Python and HTML Frontend in Heroku Server





# Launch Readiness

## Key Milestones

*Identify any relevant milestones (e.g., a Dogfood or Beta milestone) that people should know about.*

| **Date** | **Milestone** | **Description** |
| --- | --- | --- |
| Mon, Ags 21 | PRD | Domain knowledge learning, High level approach and proposed solutions |
| Mon, Sept 5 | Feature Update | EDA, Pre-processing, Feature Selection Done. Data ready for modeling |
| Mon, Sept 19 | Modeling | Expected metric result: ROC > 90%, choose the best model for deployment   * Logistic regression (classical Machine Learning) as a baseline classifier and interpretable models. * Ensemble of algorithms : RF, XGBoost, Catboost * Deep learning architecture Modeling |
| Mon, Sept 26 | Backend Serving Ready | Backend and model image is ready to deploy |
| Mon, Oct 26 | Frontend Serving Ready | GUI frontend is ready to deploy |

## Artifacts

*Put all the details and models of your milestone*

| **Artifacts** | **Where to check?** |
| --- | --- |
| Dataset Final | <https://www.kaggle.com/datasets/blastchar/telco-customer-churn> |
| Project Milestone | <https://github.com/baypsil/Telco_Churn> |
| App | <https://telco-churn-prediction-pacmann.herokuapp.com/> |

# References

<https://github.com/treselle-systems/customer_churn_analysis>