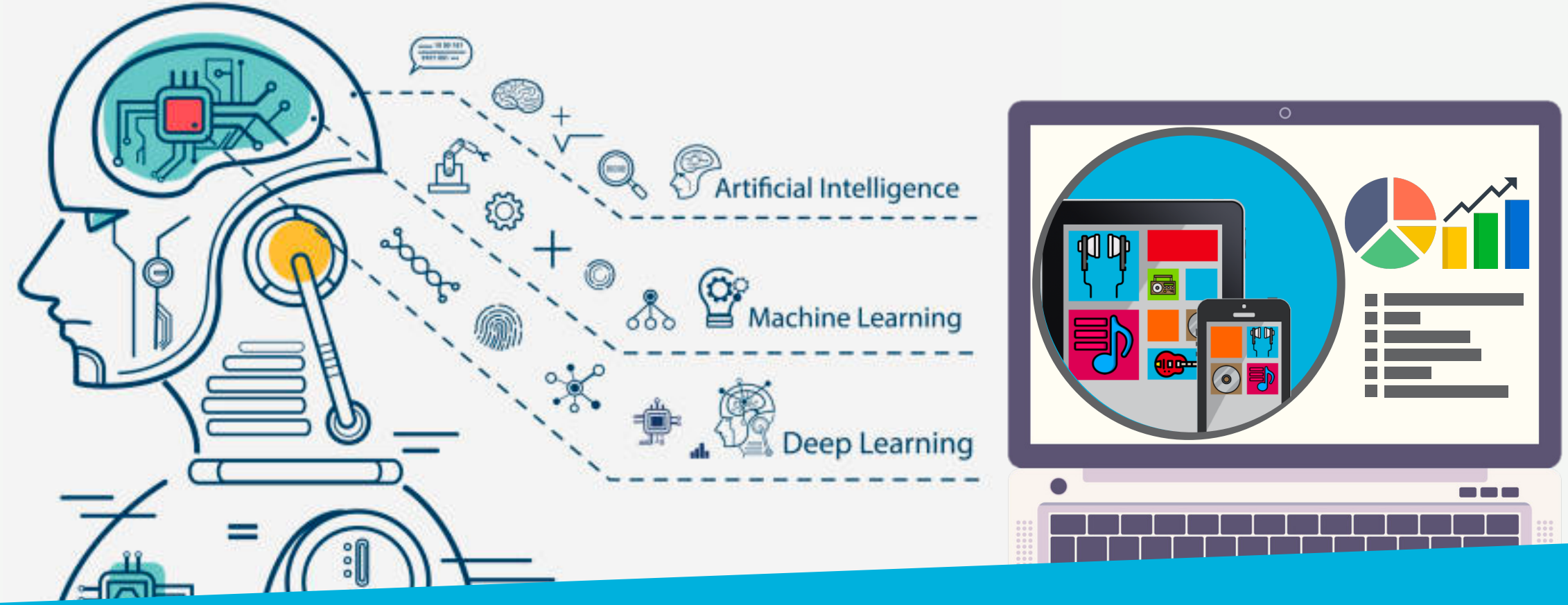
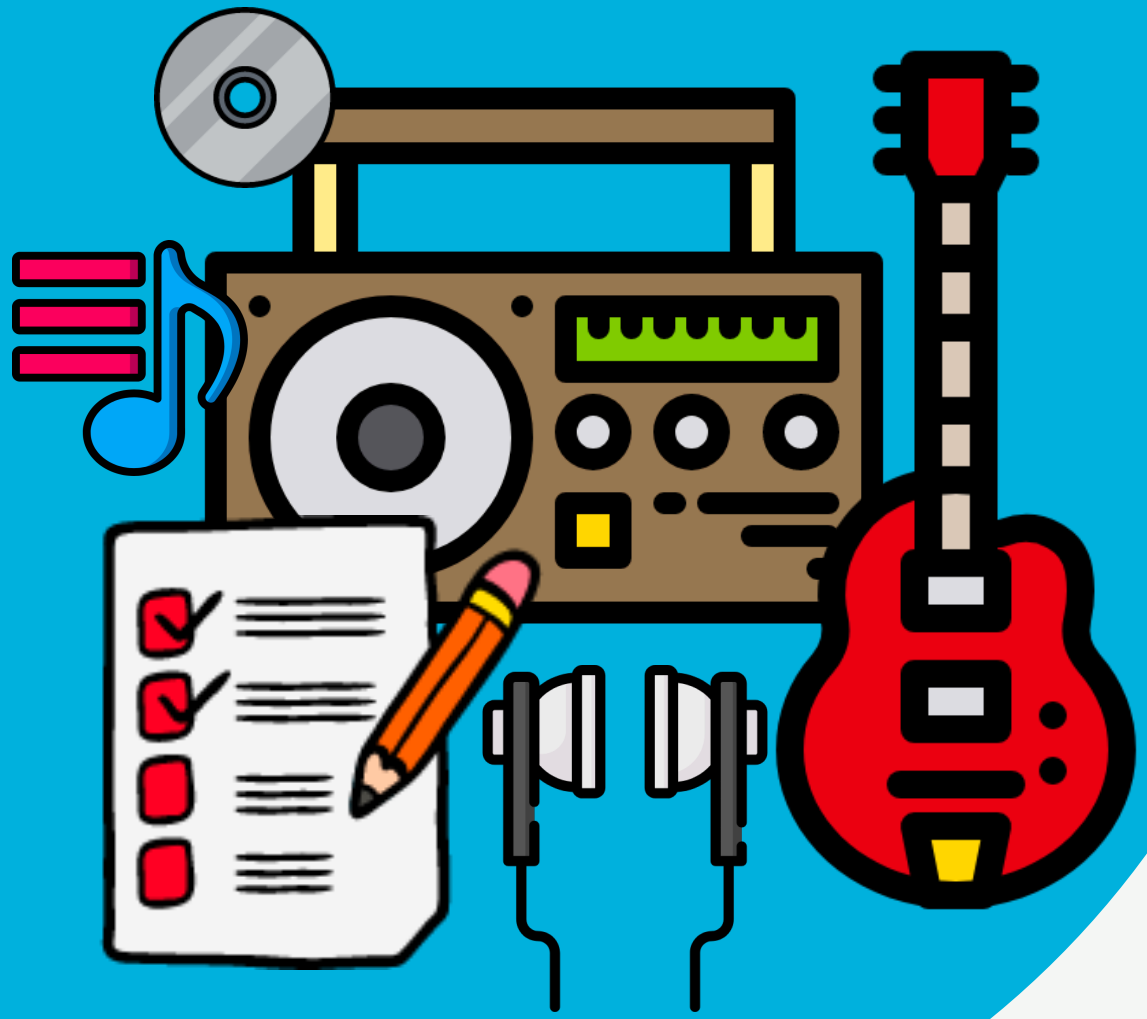
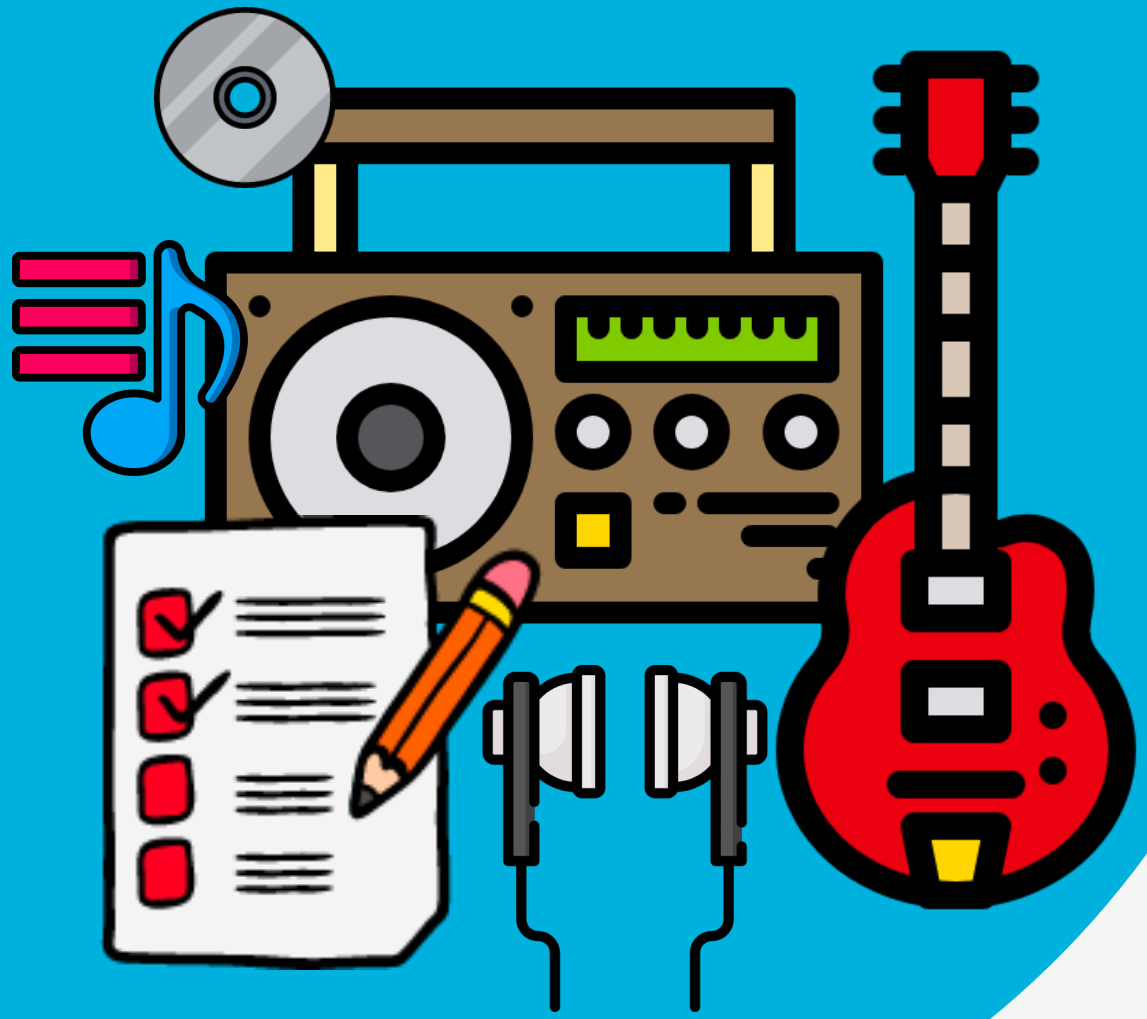


KKBox: Predicting Customer Behavior for Music Streaming Services



Agenda

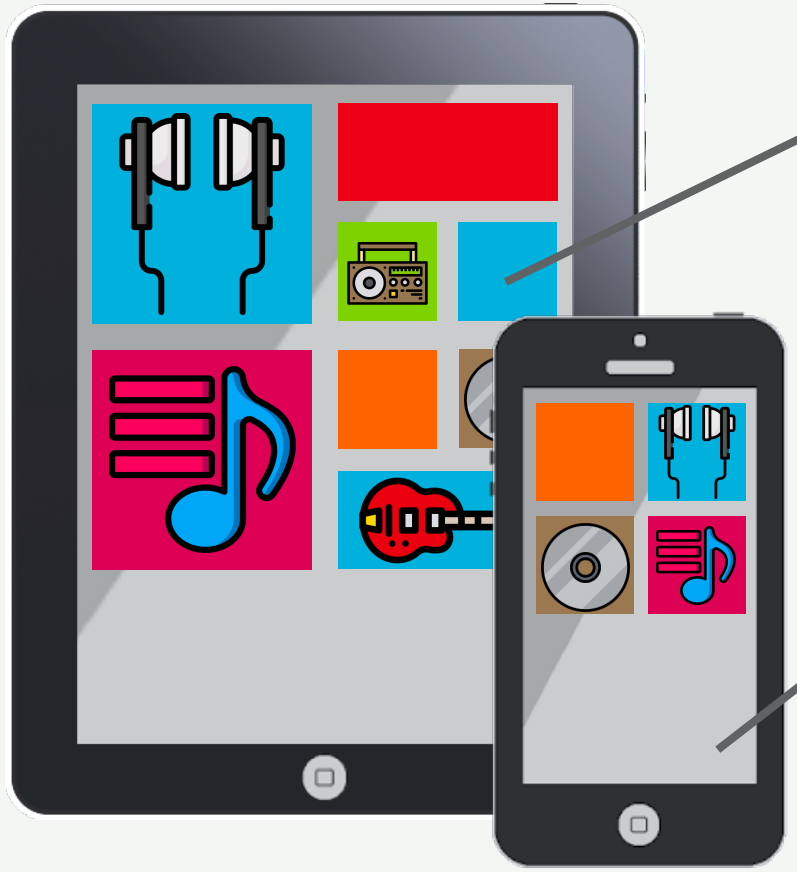
- Goals & Objectives
- Problem Statement
- Model Description
- Results & Analysis
- Further Proposals



Agenda

- Goals & Objectives
- Problem Statement
- Model Description
- Results & Analysis
- Further Proposals

Goals & Objectives

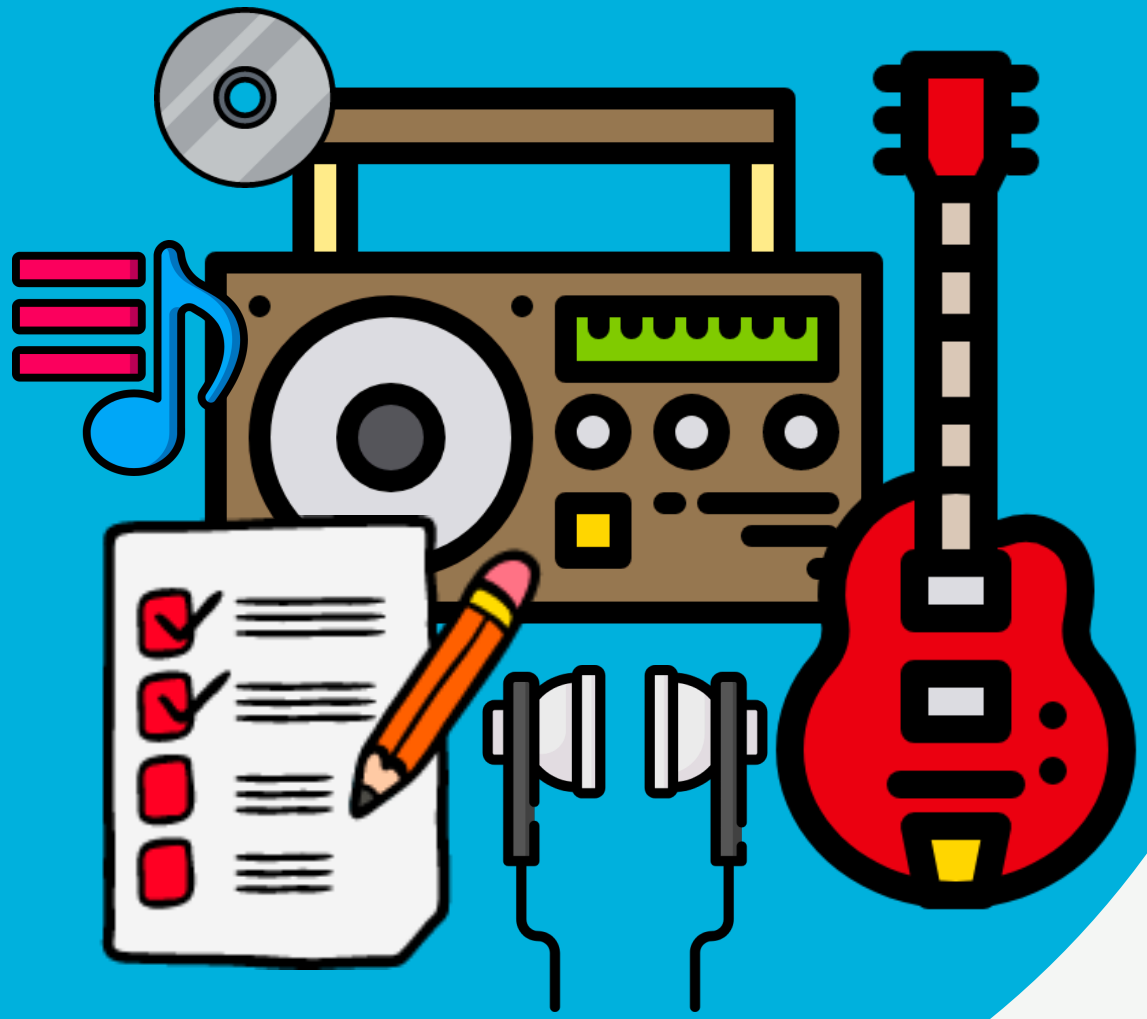


GOAL

本プロジェクトの最終ゴールは、
音楽ストリーミングサービスの
会員数を増やすことである。

OBJECTIVE

上記ゴールのために、データ分析を通し、
サービス離反顧客の判別予測をし、離反確率の
高い顧客へ集中的にマーケティングが出来る土台
を作ることが本プロジェクトの目的となっている。
具体的には、顧客の離反行動について、
一時的な離反なのか**永続的な離反**なのか
を予測するモデルを構築する。



Agenda

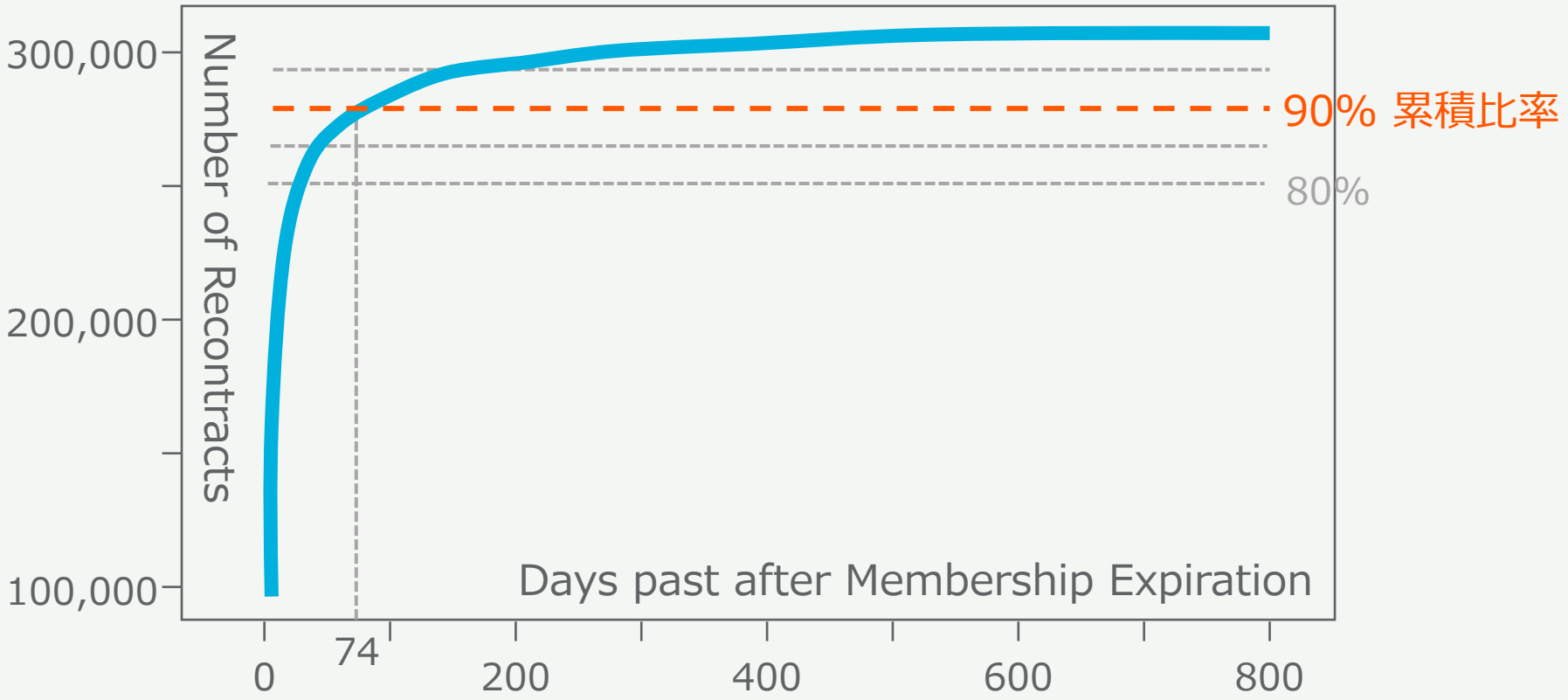
- Goals & Objectives
- Problem Statement
- Model Description
- Results & Analysis
- Further Proposals

Problem Statement

Transaction ID	Account ID	Transaction Date	Membership Expired Date	Expired Days
23	1aefooaif66hK	2015/2/3	2015/3/3	2
24	4245gjroqklaaD	2015/4/30	2016/4/30	376
⋮	⋮	⋮	⋮	⋮
309423	887HHKUH9	2017/3/15	2017/3/22	62

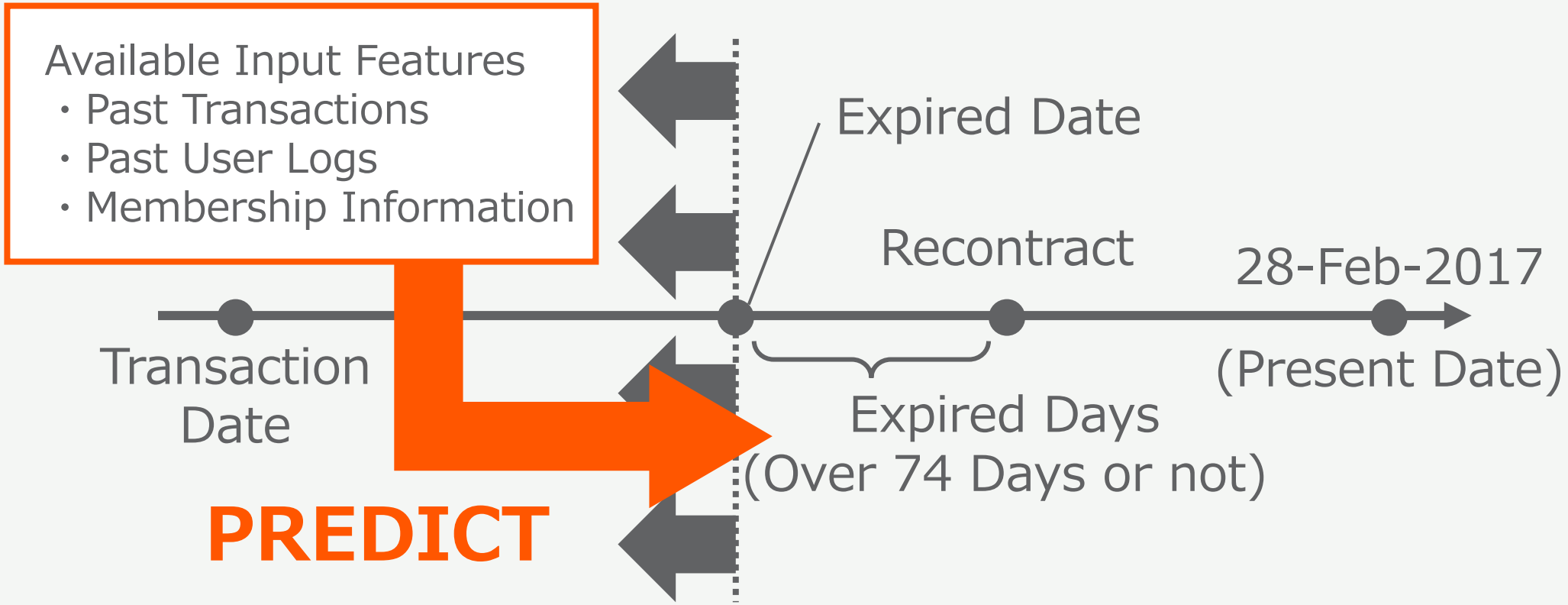
音楽ストリーミングサービスにおけるユーザーアカウントが失効したレコードの中で、
Expired Daysがある値を超えると永続的な離反と見なす

Expired Days Threshold



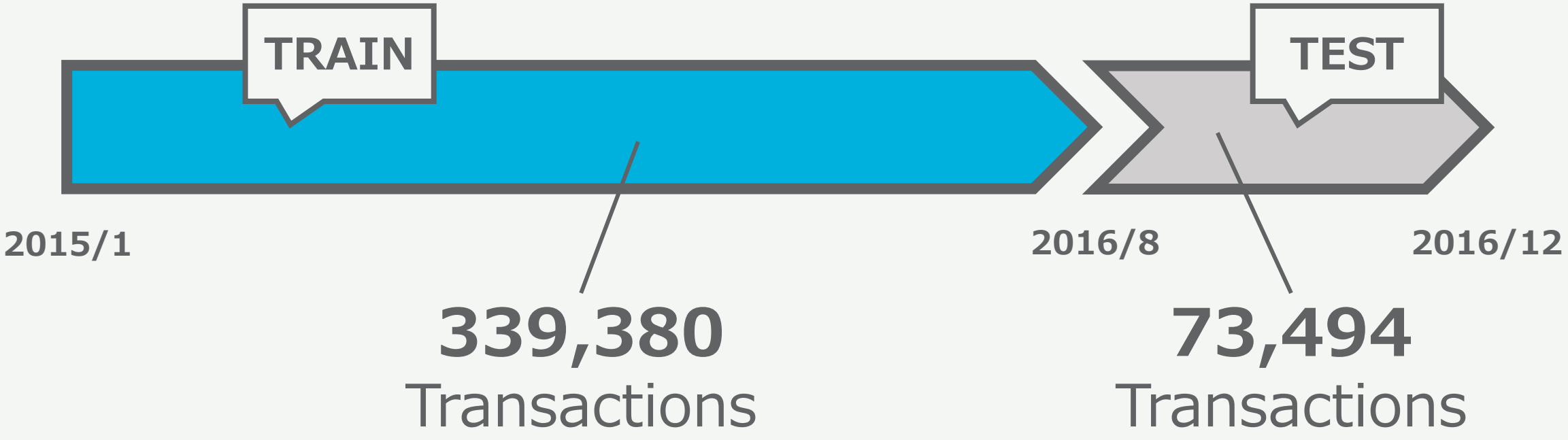
アカウントが失効した際に、一時的な離反か永続的な離反かを区切るしきい値には
再契約者のパレート図で傾向が変化した**74日**を採用した

Available Features



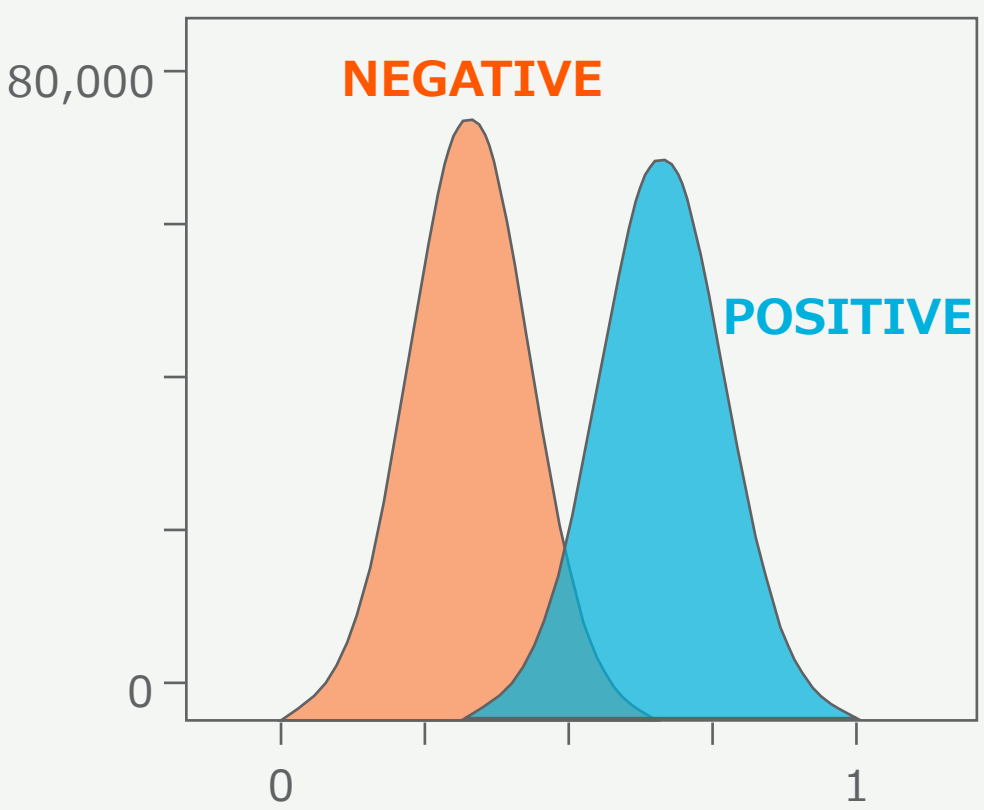
使用出来る特徴量は、過去に起きたTransaction、
過去のサービス利用ログ、アカウント情報の3点のみ

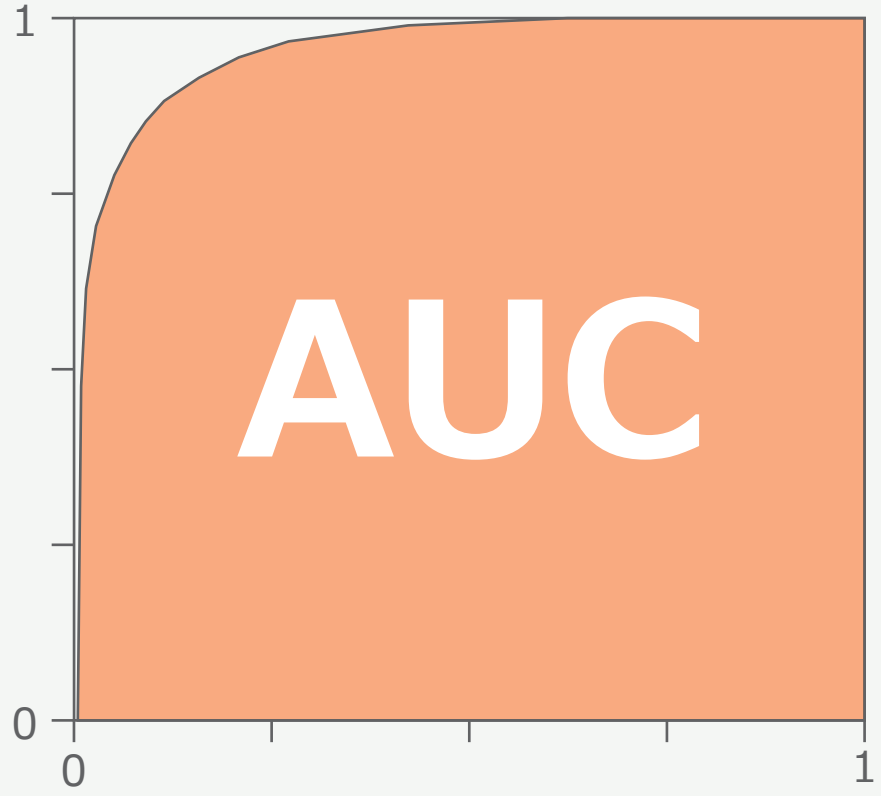
Evaluation Method



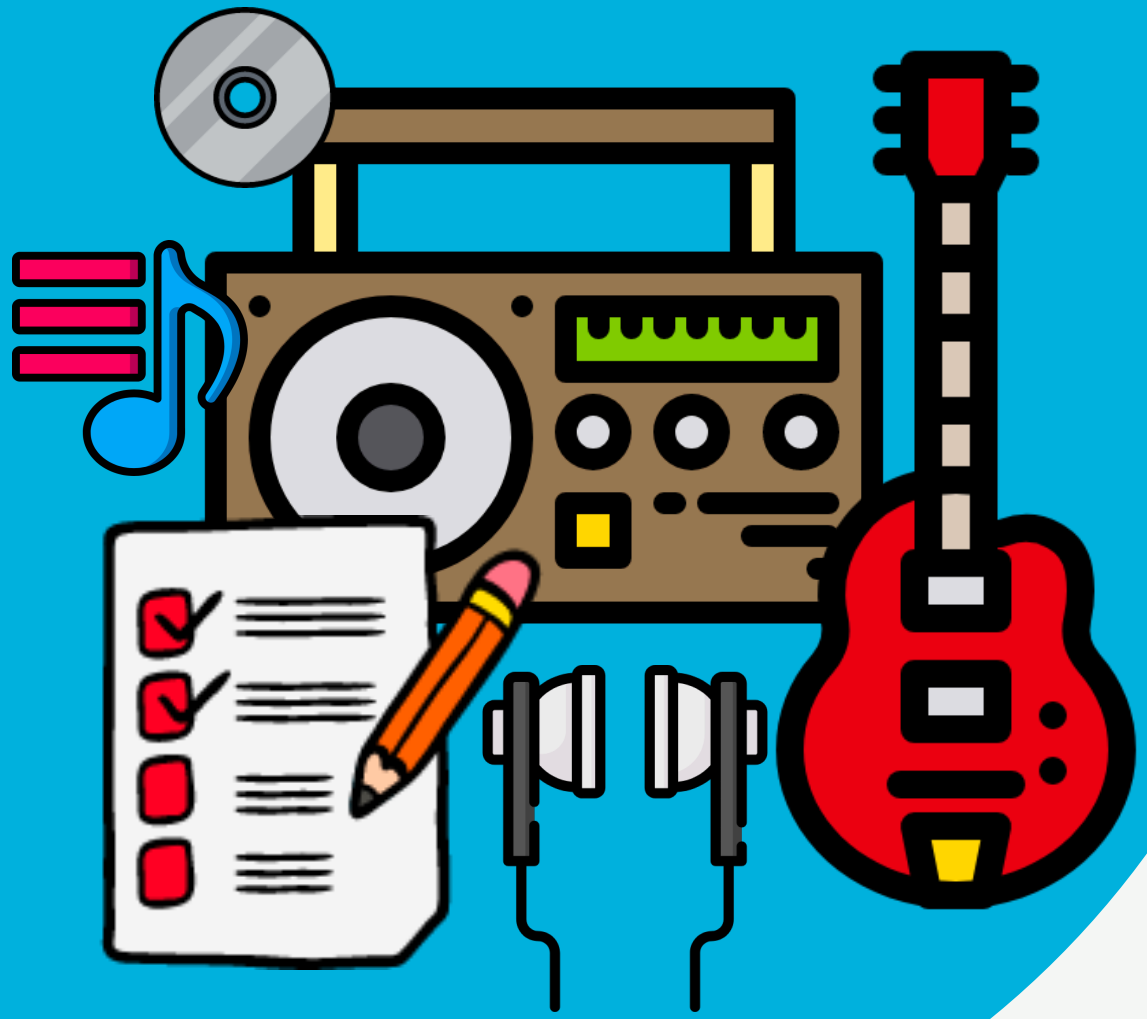
実運用になるべく近づけるために、
時間軸が後ろのものを評価用テストデータとした

ROC AUC Score





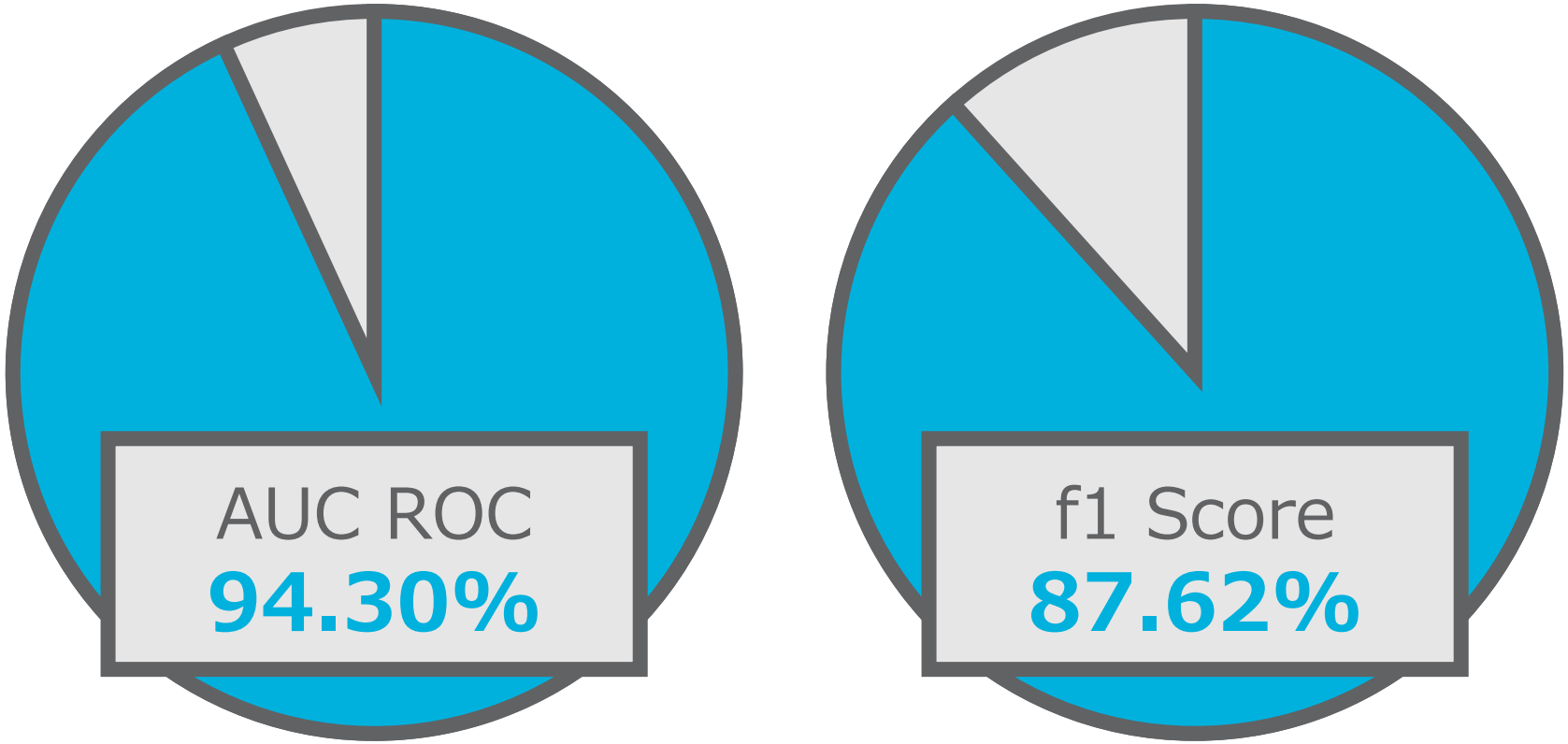
評価の指標として、データの特性に大きくよらずモデルの優位性を
評価出来る、ROC曲線のAUC(Area Under the Curve)を採用



Agenda

- Goals & Objectives
- Problem Statement
- Model Description
- Results & Analysis
- Further Proposals

Results : Scores



テストデータに対するスコアは上記の通り。

Model Description

LIGHTGBM　　DEEP LEARNING

上記２種類のモデルを組み合わせて離反顧客の判別予測を実施した。

LIGHTGBM

Model Description

Basic Description

・勾配Boosting系列の機械学習モデル
・予測精度を競う世界大会kaggleでは今最も使用されているモデル

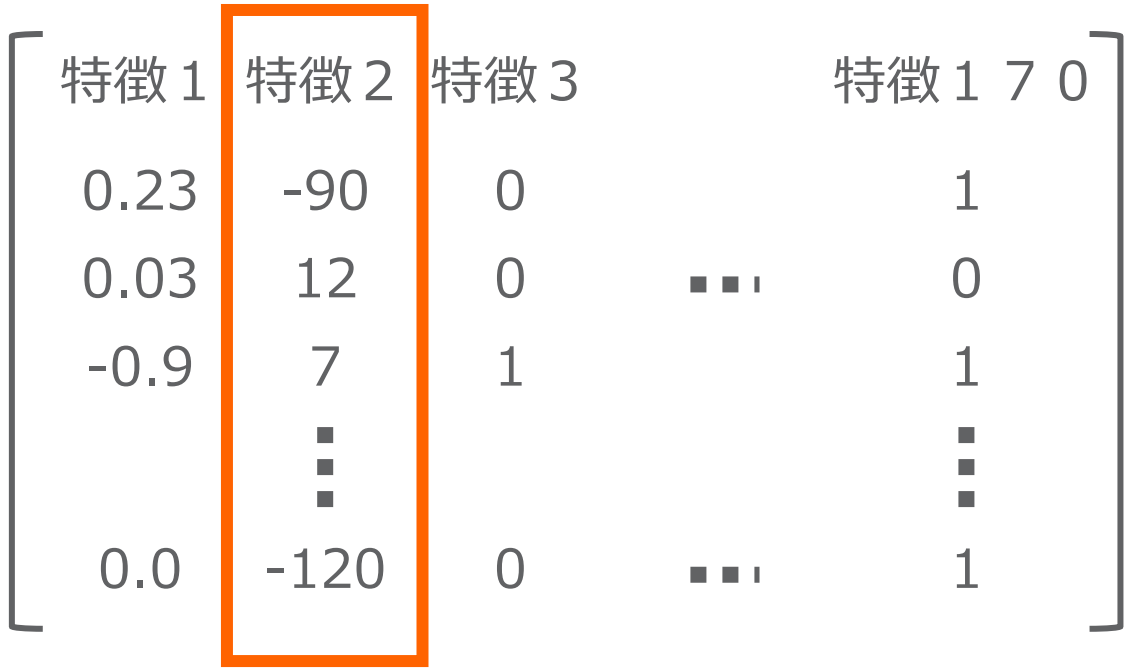
Algorithm

・決定木モデルを構築し、その決定木モデルの予測値と正解の差を後続の決定木が予測する
・最終的な特徴量は26個となった

LIGHTGBM

Feature Selection

SHUFFLE



BASELINE SCORE − **SHUFFLED SCORE**

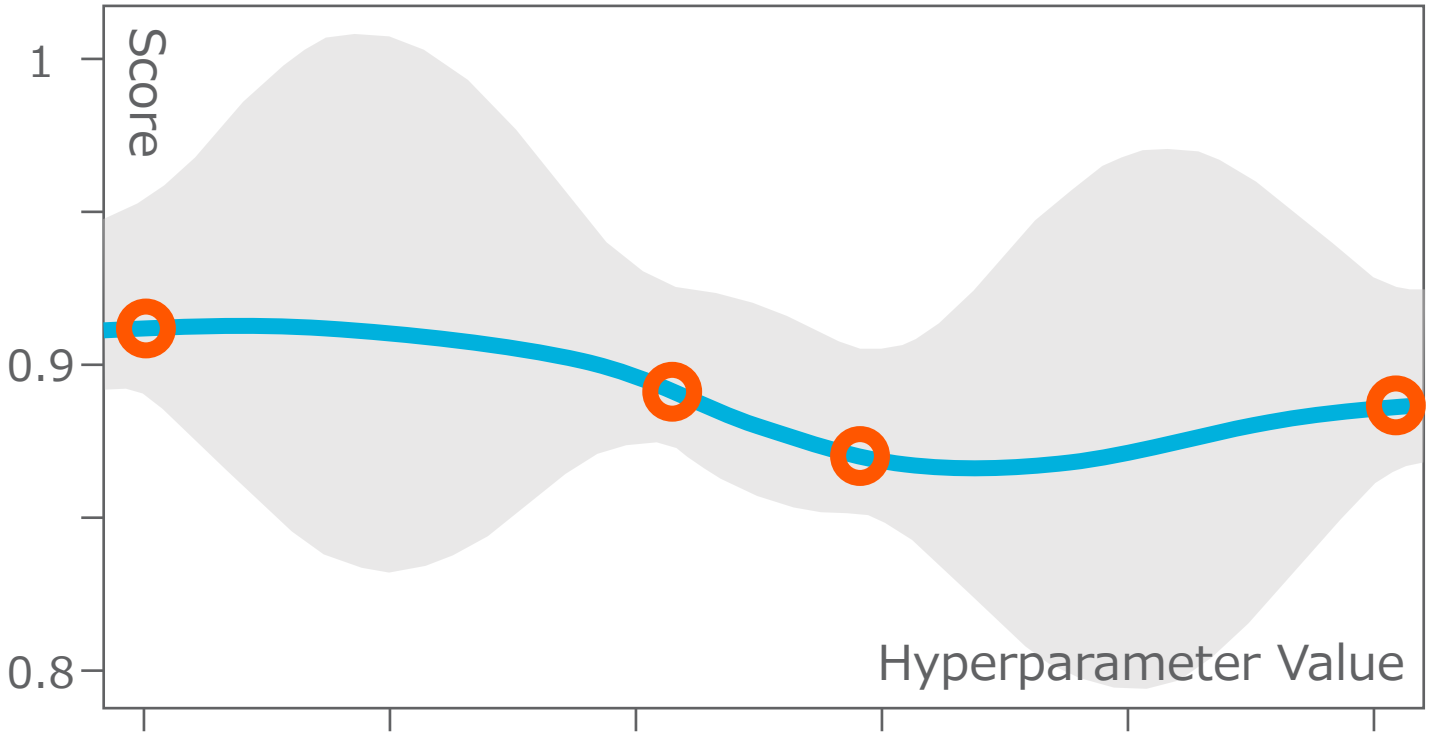
= Permutation Importance

特徴量の選択には、ラッパー法(※APPENDIX#1)ながら計算時間が比較的短い

Permutation Importanceによる特徴量選択を実施

LIGHTGBM

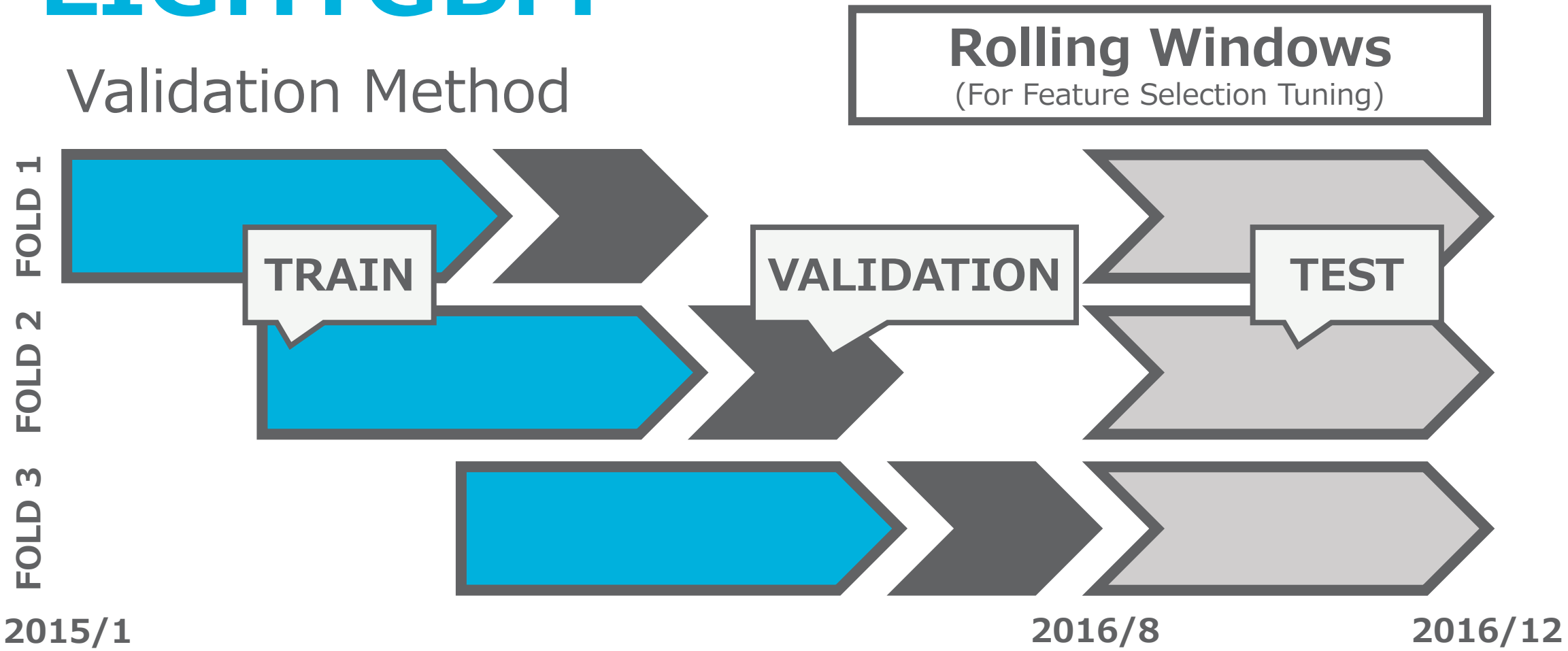
Hyperparameter Tuning



モデルのパラメータ調整には、**Bayesian Optimization**と呼ばれる、
ベイズの定理をベースとした最適パラメータの探索アルゴリズムを使用

LIGHTGBM

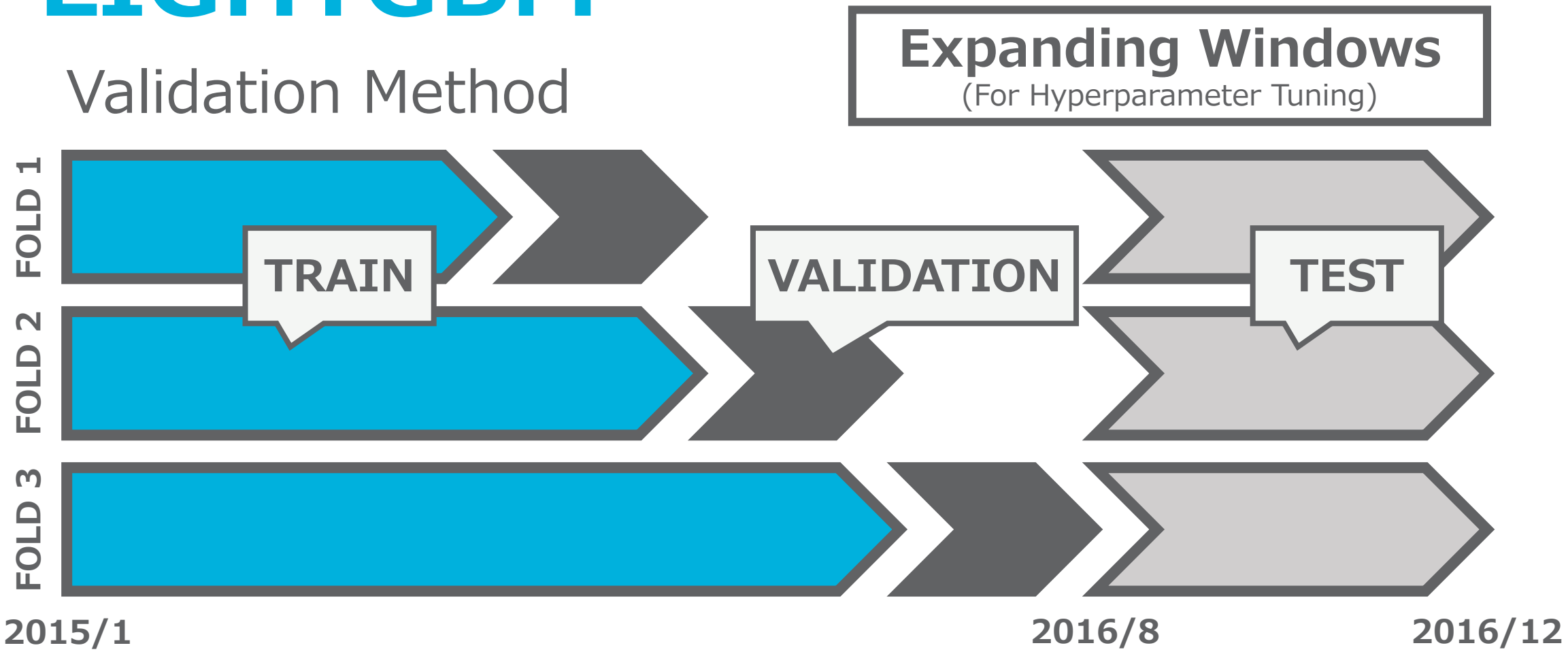
Validation Method



モデルの検証は、過去データから未来データを予測する形を崩さずに、
幅広い時間軸で検証出来るように**TimeSeriesSplit方式**の交差検証法を実施

LIGHTGBM

Validation Method



モデルの検証は、過去データから未来データを予測する形を崩さずに、
幅広い時間軸で検証出来るように**TimeSeriesSplit方式**の交差検証法を実施

DEEP LEARNING

Model Description

Basic Description

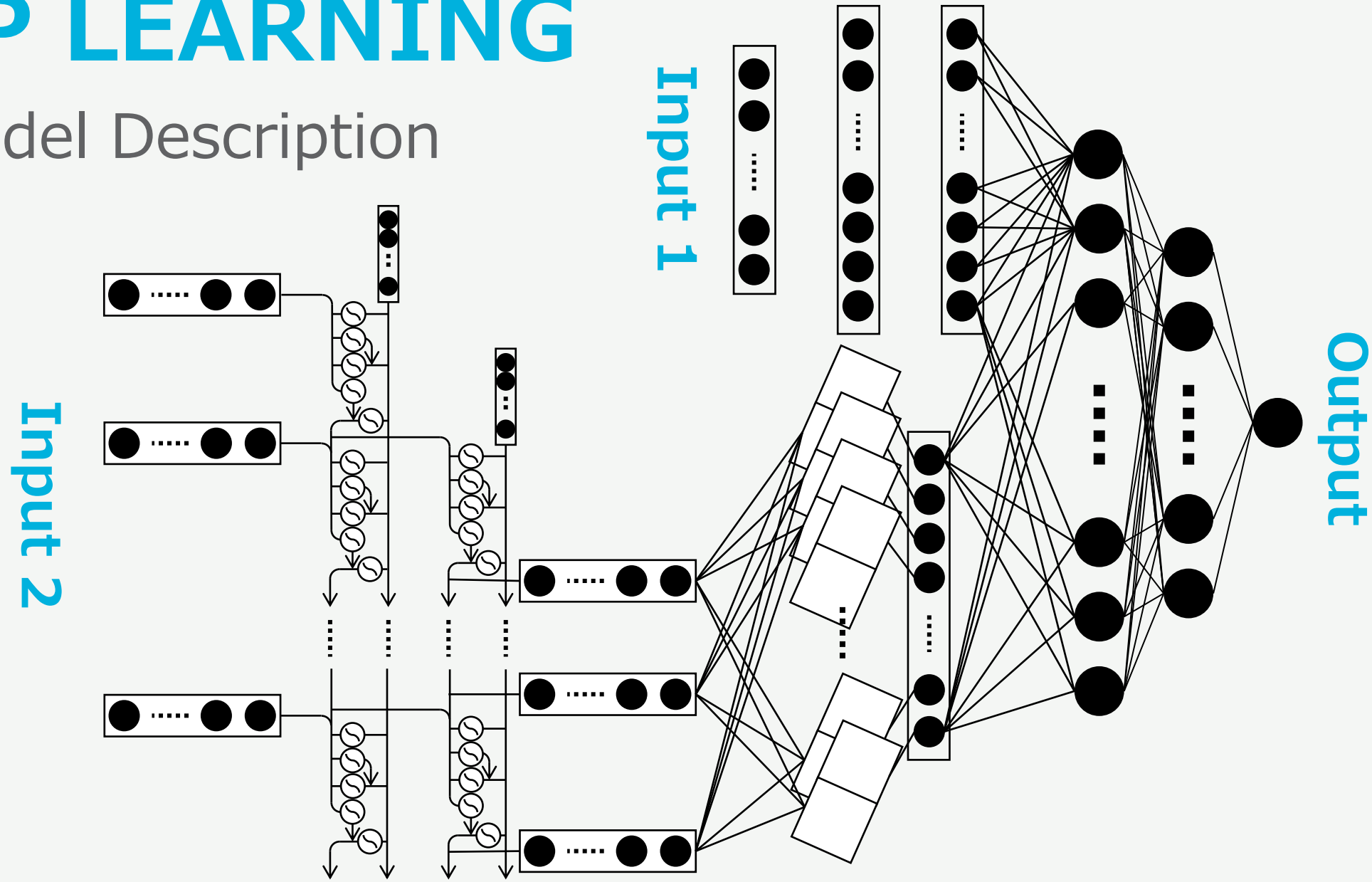
・時系列データを扱うのに長けている
LSTMがベース構造
・機械学習モデルよりもミクロな特徴
を掴むことが出来る

Algorithm

・LSTM用のサービス利用ログの
インプットと、機械学習モデルと
同じ特徴量を学習するインプットの
2種類インプットを作成する

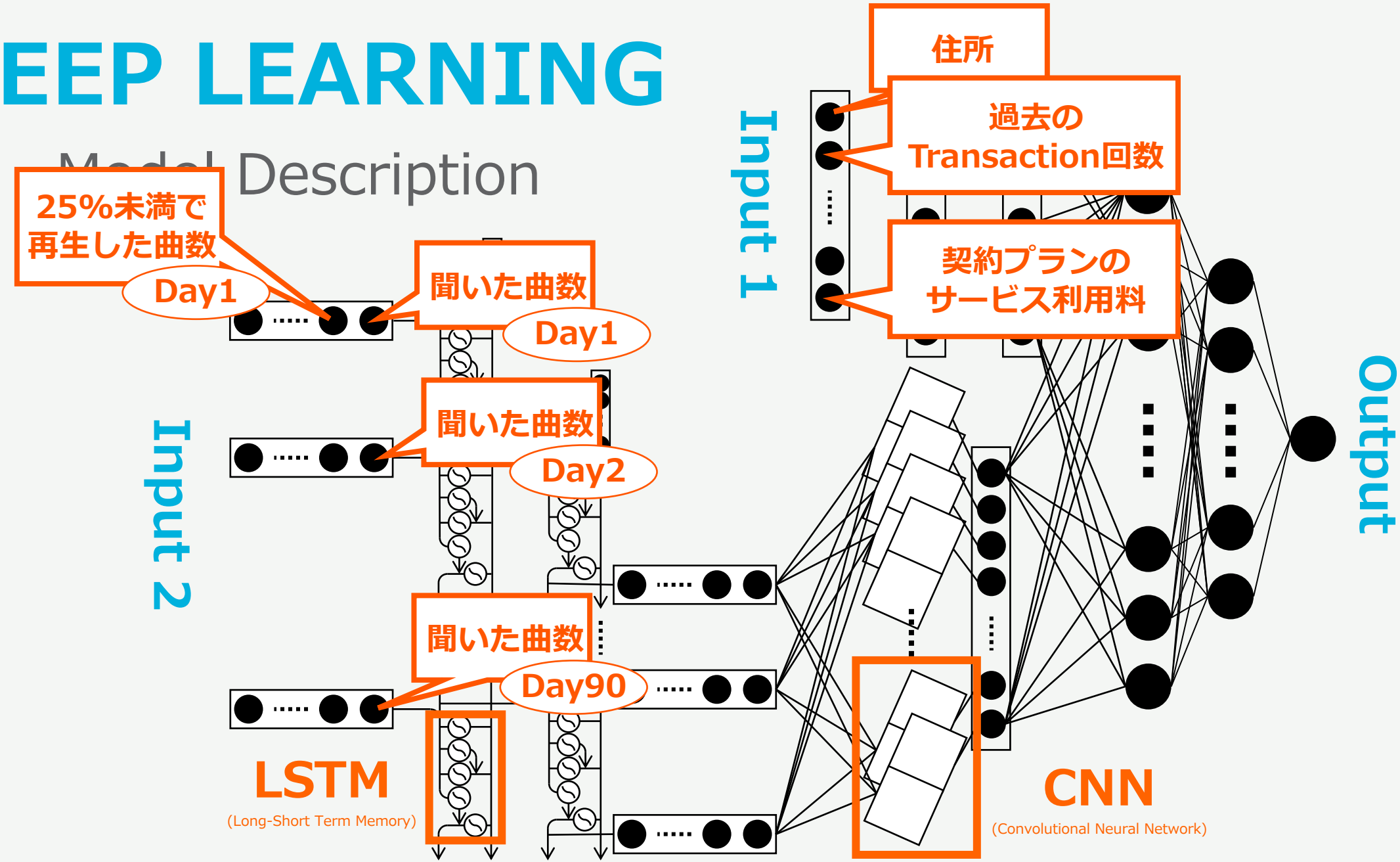
DEEP LEARNING

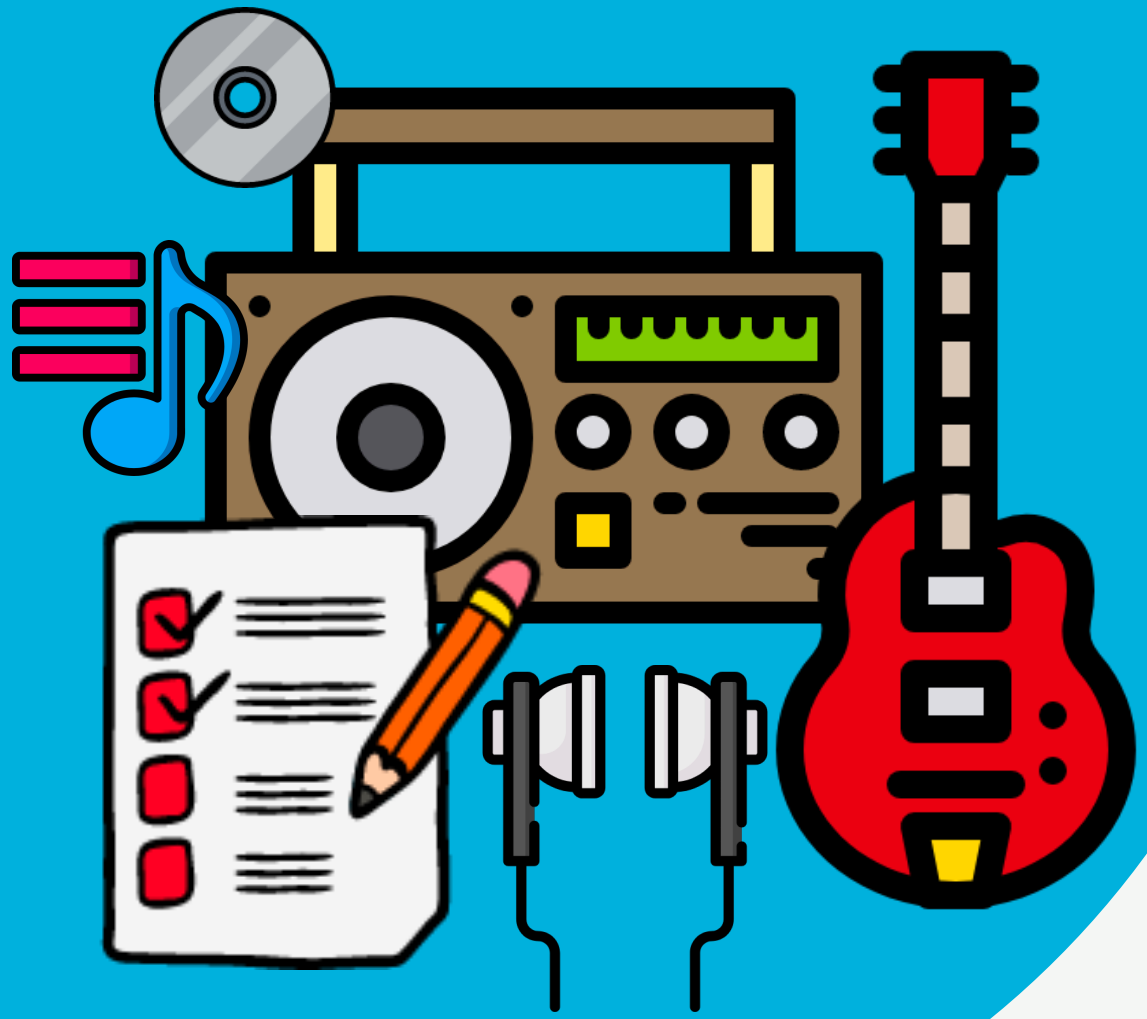
Model Description



DEEP LEARNING

Model Description

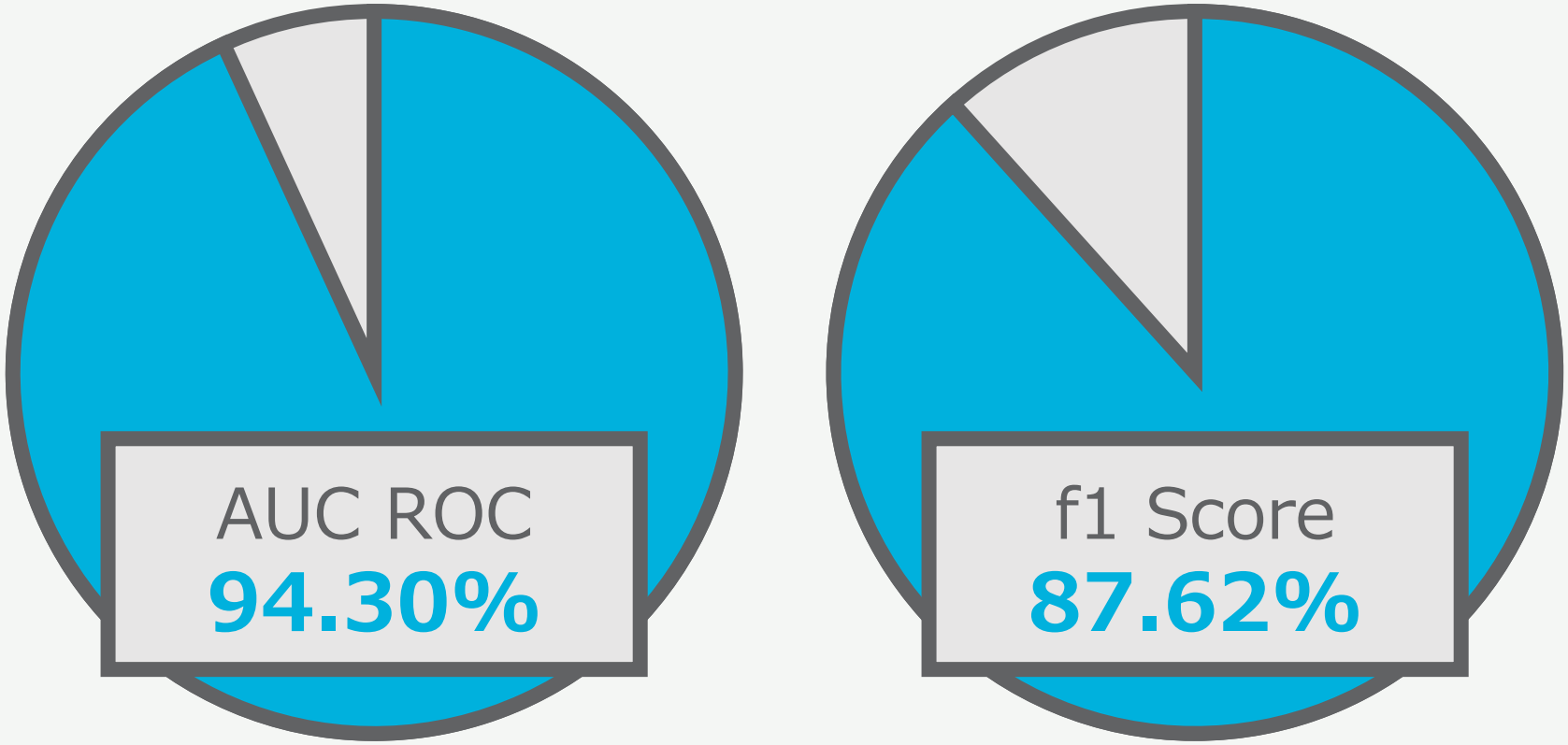




Agenda

- Goals & Objectives
- Problem Statement
- Model Description
- Results & Analysis
- Further Proposals

Results : Scores



テストデータに対するスコアは上記の通り。

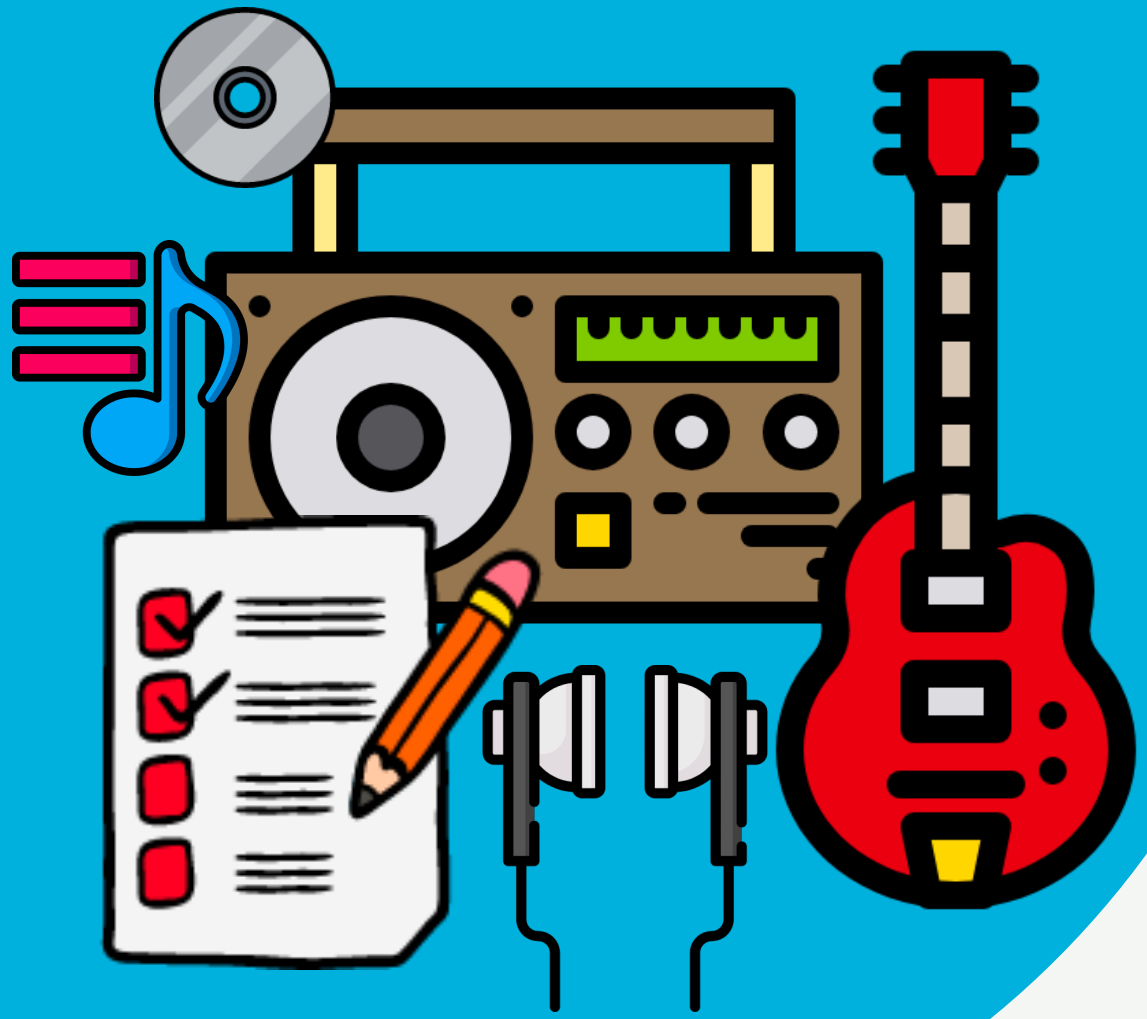
Results：Confusion Matrix

	一時的な離反	永続的な離反	
一時的な離反（予想）	**31,694** Transactions	**3,763** Transactions	Recall：**85.68%**
永続的な離反（予想）	**5,445** Transactions	**32,592** Transactions	
	Precision：**89.64%**		f1 Score：**87.62%**

Analysis : Useful Features

num_past_expired_transactions_log	-0.091358	過去の離反回数
num_logs_last3months	-0.042701	過去3ヶ月のサービス利用ログ
is_auto_renew	-0.024783	自動更新かどうか
payment_plan_days_log	-0.015758	契約プランのサービス利用料
mean_expired_days_log	-0.011396	過去の離反日数平均
actual_amount_paid_log	-0.006963	支払ったサービス利用料
city_1	-0.002877	住所：都市IDが#1
num_unique_last3months_log	-0.002798	過去3ヶ月の音楽を聞いた曲の種類数
sum_actual_amount_paid_log	-0.002515	過去の累計サービス利用料
payment_plan_days_categorized_30	-0.002074	契約プランの日数：30日
total_secs3months	-0.002030	過去3ヶ月の音楽を聞いた秒数合計
total_sec_skew3months	-0.001777	過去3ヶ月の音楽を聞いた秒数の尖度

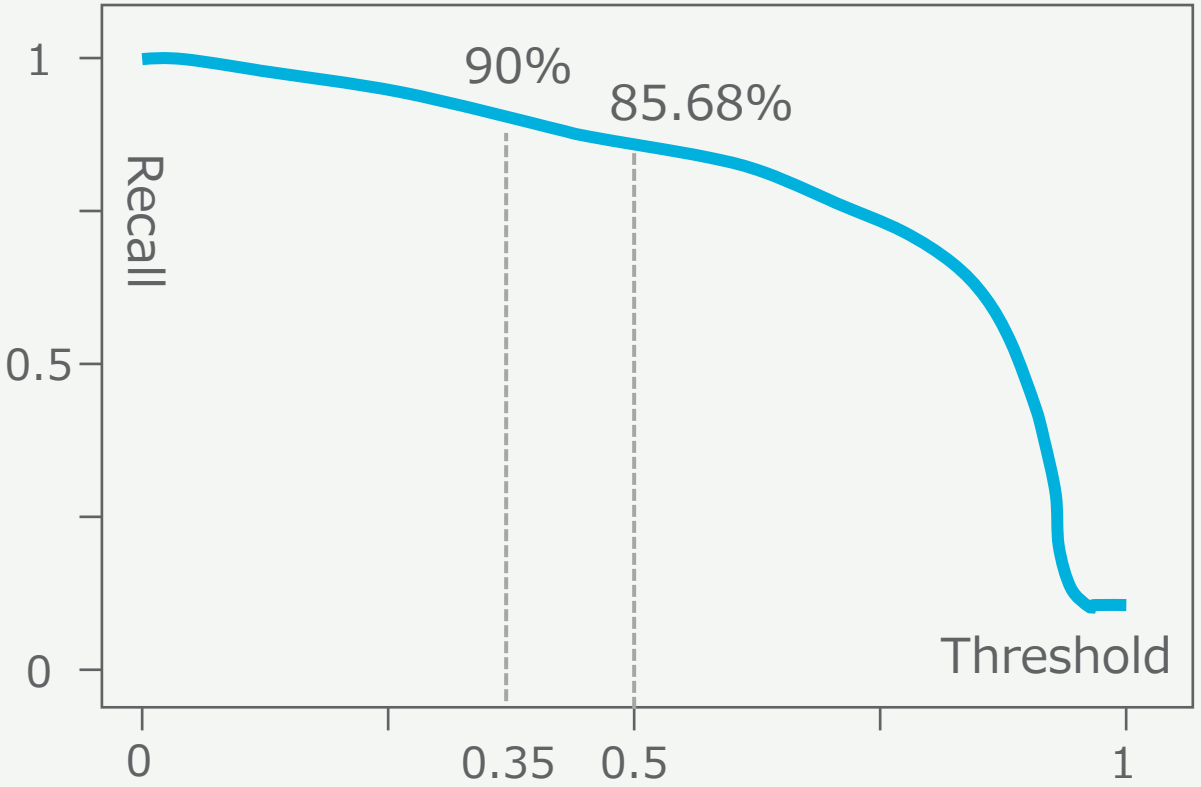
Permutation Importanceの上位26個の特徴を使用した
上記は上位12個の特徴である



Agenda

- Goals & Objectives
- Problem Statement
- Model Description
- Results & Analysis
- Further Proposals

Further Proposals



モデルのしきい値を0.35にすると、Recallが90%となり、
永続的な離反顧客をより高確率で当てられるようになる

Further Proposals

LIGHTGBM

AUC ROC 94.16

◎ モデル運用が低コスト

特徴量が26個と少なく、時間軸の変化に
よらない汎用的な特徴を使用しているため、
モデルの再学習が頻繁に必要としない

◎ 予測結果への説明力が高い

比較的単純な機械学習モデルなため、モデルの
予測結果に対する説明力が強い

DEEP LEARNING

AUC ROC 94.29

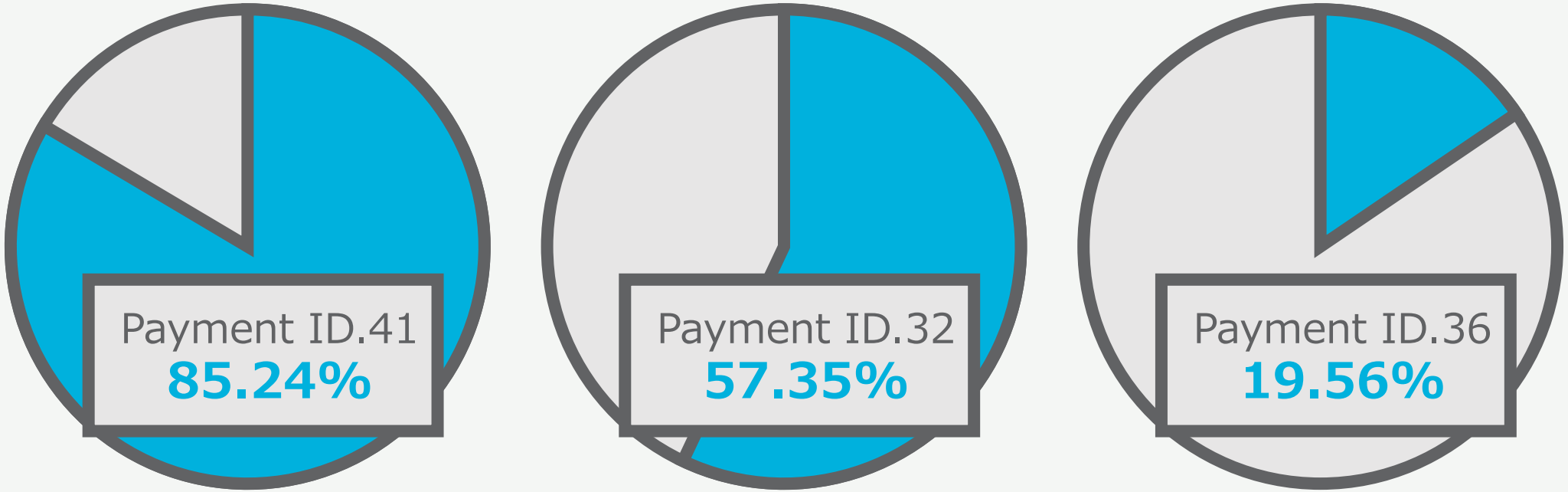
◎ 予測精度が高い

離反顧客を正確に予測しているため、
費用対効果が高いマーケティングが
可能

モデルの運用工数・運用コストを抑えたい場合、LightGBMモデルのみ使用
モデルの精度を高め、マーケティングコストを抑えたい場合、両モデルを使用

Further Proposals

NEGATIVE



POSITIVE

Payment Methodによって正例、負例の割合に大きな差があったため、
Payment Methodの選択によってクーポンを配ることも離反を回避出来るかもしれない

APPENDIX

APPENDIX #1

	フィルタ法	ラッパー法	埋め込み法
計算時間	◎	△	○
予測精度	△	◎	○
	相関などの統計量を使って選択する方法。一般的に高速だが、精度の点で難がある。	モデルの学習と特徴選択を何度も繰り返すことでベストな組み合わせを見つける方法。時間はかかるが、予測精度は高い。	モデルの学習と同時に使用する変数を学習する方法。計算時間と精度のバランスが良い。