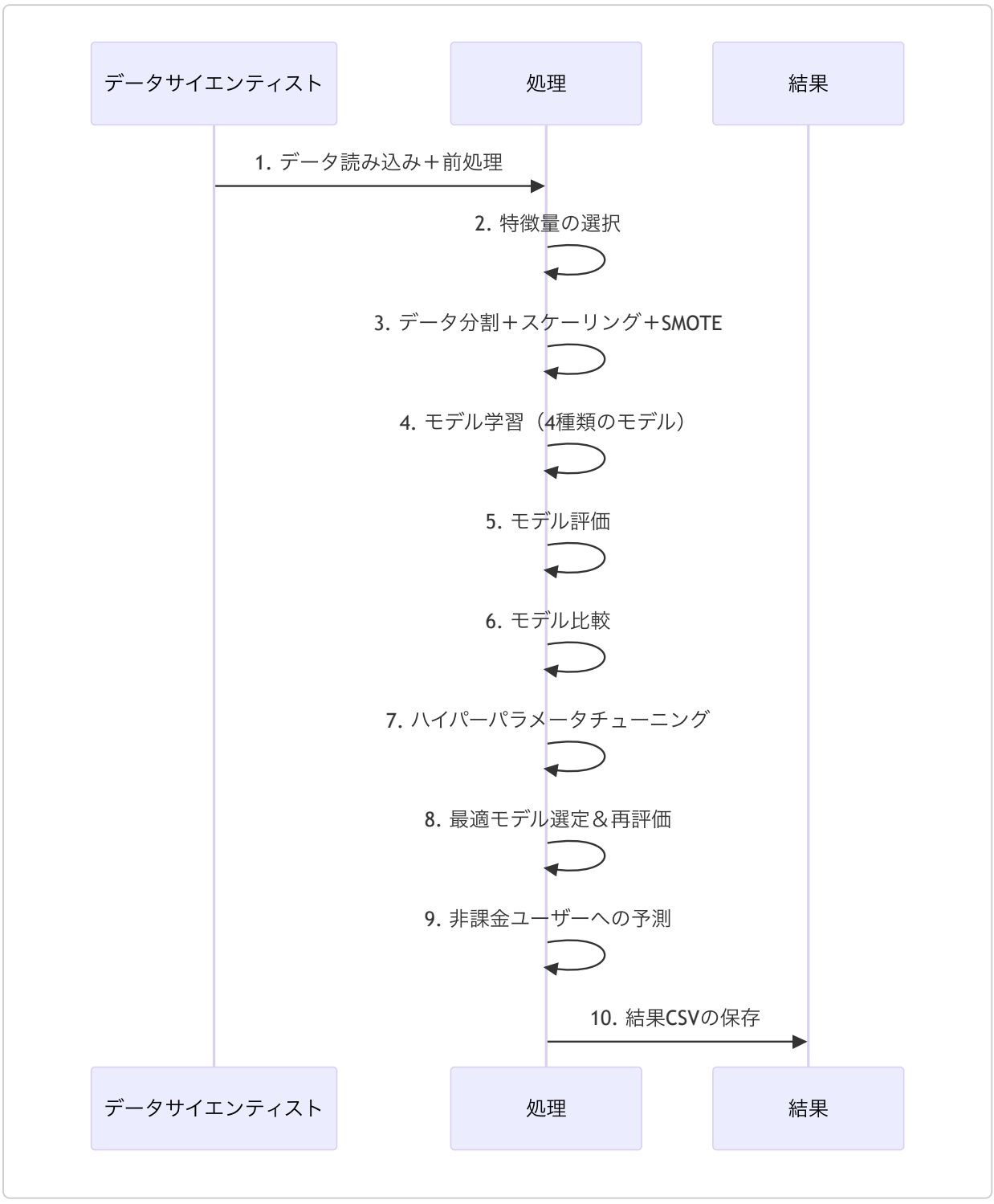
課金化予測モデル 概要

1. 処理フロー



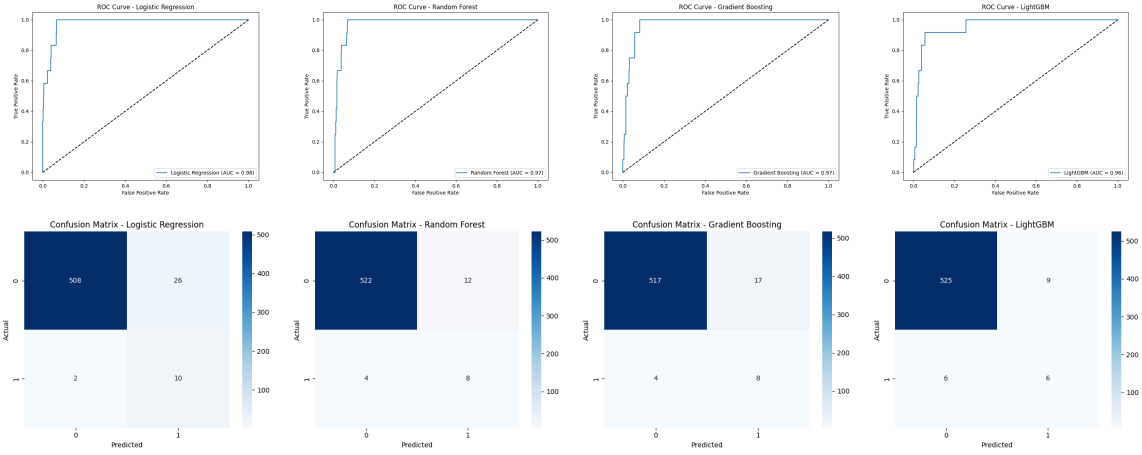
2. 処理フローの解説

1. **データ読み込み + 前処理**
   DynamoDB からエクスポートされた JSON ファイルを読み込み、 **欠損値の補完** や **不要な列の削除** を行います。また、 **status_code=2** を課金ユーザーとして扱うなど、予測に必要な前処理を実施します。

2. **特徴量の選択**
   画像ラベル数 (`image_label_count`)、年齢 (`age`)、経過時間 (`updated_time_diff`)、 **テキストから抽出した「依存度」「経済的余裕」スコア** など、モデルに取り込みたい特徴量を抽出します。

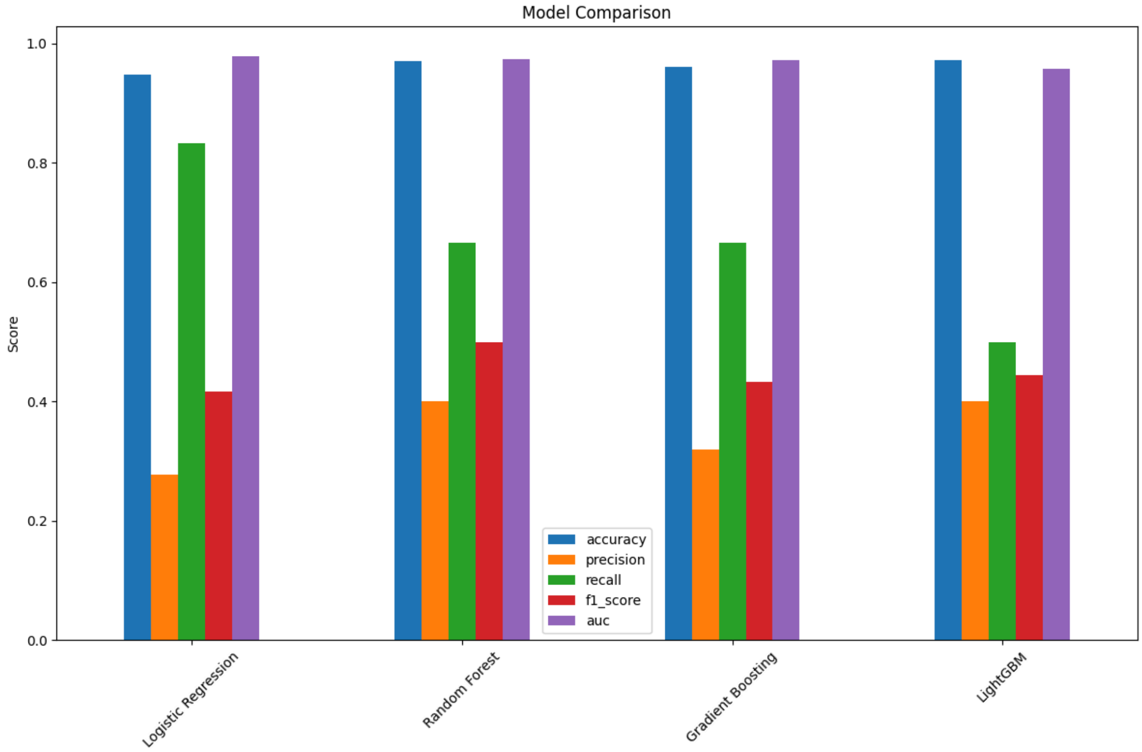
3. **データ分割 + スケーリング + SMOTE**
   データを「学習用」と「テスト用」に分割し、数値を **標準化** します。さらに、課金ユーザーが少ない不均衡データに対して、 **SMOTE** （少数クラスのオーバーサンプリング）を適用し、学習バランスを改善します。

4. **モデル学習（4種類）**
   ロジスティック回帰・ランダムフォレスト・勾配ブースティング・LightGBM の 4 種を同時に学習させます。



5. **モデル評価**
   テストデータに対して、精度（Accuracy）、適合率（Precision）、再現率（Recall）、F1スコア、AUCなど、多面的な指標で性能をチェックします。

6. **モデル比較**
   上記の評価指標を一覧化し、どのモデルが最も優れているか可視化します。



7. **ハイパーパラメータチューニング**
   各モデルのパラメータ（例：木の深さ、学習率等）を細かく試し、 **より高い精度** を出せる設定を自動的に探索します。

8. **最適モデル選定 & 再評価**
   チューニングで最良の結果を出したモデル（例：LightGBM）を最終採用し、再度評価して最終的なスコアを確定します。

9. **非課金ユーザーへの予測**
   「現時点で課金していないユーザー」の課金化率（課金する確率）を計算し、 **優先的にアプローチすべき顧客** を選別できます。

10. **結果CSVの保存**
    最終的に、課金化率が閾値以上のユーザーをCSVファイルなどで一覧出力し、マーケティング施策などに活用します。

non_paying_users_with_probabilities

lineuid	nickname	conversion_probability
Uf56f93feb33717d184b2973ddeac6605	☆ayumi☆	0.8293168043095900
U380bbcd939f645ce8464bbdf4cb8164b	YUKO🎵	0.3986775670730890
Ubd16b47ef94309a061140a346130db33	takacoค(´ω｀ค)ว	0.3789688470641310
U7db43efb10a528762e7720660db65285	☆金子真夕子☆	0.37640218682910700
U725c0f376c68ff541e60eed74bb9a45a	Unknown	0.23923476399763600

non_paying_users_with_probabilities

lineuid	nickname	conversion_probability
Ubd16b47ef94309a061140a346130db33	takacoค(´ω｀ค)ว	0.6145089830105640
U725c0f376c68ff541e60eed74bb9a45a	Unknown	0.579832531752356
U7db43efb10a528762e7720660db65285	☆金子真夕子☆	0.30407047145981600
Uf56f93feb33717d184b2973ddeac6605	☆ayumi☆	0.24201998343658200
Uece21687fa3410a840b29d4dea22fd81	渡邉 未来	0.23594215476292000
U13400b658849576926f3cb768ed1a5c7	川久保	0.22689671746749400
Uf96b0ba81ec6f0f7fd490a102610470b	椿　ミヨシ	0.19651732705306700

non_paying_users_with_probabilities

lineuid	nickname	conversion_probability
U13400b658849576926f3cb768ed1a5c7	川久保	0.9664445448887020
Uece21687fa3410a840b29d4dea22fd81	渡邉 未来	0.8873170187717840
U8fff8d109460872192e8c15650e3072f	さかいひろみ	0.873483662511047
Ubafdc323d37c1b93e6923a622269453a	Kyoko✨	0.81635486658009
Uf96b0ba81ec6f0f7fd490a102610470b	椿　ミヨシ	0.8107312505056200
U8432cdef5ed5d7428e5bc98418df3cee	ひらしま　なおみ	0.7859478896371190
Uf56f93feb33717d184b2973ddeac6605	☆ayumi☆	0.766030529801942
Ubd16b47ef94309a061140a346130db33	takacoค(´ω｀ค)ว	0.746291527769617
Uf656f6d1f0005ef3d215e1f063865c21	Tuula	0.7404914000086750
Ue509bb3e05bad5c67f9b5bdfb2b51a43	megu	0.6707581589827350
Ua4c4b53d79f82d935cd1756ca6eb8a76	だだだいき	0.6560959786824210
U87331f04ee999636be15a7e5a72b91e2	あき	0.6511848434369090
U7de93a654a3a6034a71fa90022d004a2	大久保由美子	0.626161139701395
U2bdc116e53bd4bfa64aede877ebfa739	mariko	0.5588577456941430
U3d3535cd263eae1f12751bc620c19502	kachy	0.5479929381869720
U6adddb1008902810899a017e72b71ae9	TF	0.5376979166431810
U4f381bd8393627c8b8359072500d3dea	SAORI ISHIDA	0.5247950876473730
U51908c473ca0212b92c53c7452a69614	麻由(岳ママ★雛ママ)	0.5021919929555430
Ud6d52e2b8c20f9c442d5d9723522df77	前島 真由美	0.499093731884483
U132e13b38a944926e190e973bc554286	きよ	0.477433455262815
U1f98a674327cbbdb3fc8a6c58fa9a041	Rie	0.46043063782187200
U372070ee96953e5eb1901d0844281f4a	♡YU♡	0.4456009691202740
U95d48e02f352e72e85a7582453320270	松本 智恵	0.4410131085004560
U8ed838c44c882b1b140d6a4a4e3d770c	m.yukari❤みーこ ❤	0.4387369199055690
Ucd62ab1e08d70f85f3ded5ffdec40eb5	前.川亜：紀：：.	0.43512539311319700
U79e377ba8b7b0ccf5c08b25cc50a813a	あや	0.42071563753226800
Ue133893a1d2cdd7c0354695a9cd8fe97	🍎	0.419520952353982
Ud8e3a1012615791f721972e4cfda9339	naomi.sumitani♡	0.4045784131219930
U922889d38fc55069d04d2d1518a95f9d	＊MEGU＊	0.3904324978867220

3. ログの概要説明

ログには、実際の実行結果として以下が含まれています。

- 欠損値の確認結果、カテゴリカル列の情報、最終的に使用する特徴量のリスト
- 4種類のモデル（Logistic Regression, Random Forest, Gradient Boosting, LightGBM）の評価結果（Accuracy、Precision、Recall、F1、AUCなど）
- モデル同士の比較結果
- ハイパーパラメータチューニングの最適パラメータとスコア
- 最終的に選ばれたモデル（例：LightGBM）での評価メトリクスと混同行列
- 非課金ユーザーへの予測確率を算出し、CSVへ保存する流れ

4. ビジネス上のメリット

この仕組みにより、**「課金する可能性の高いユーザー」**を効率よく抽出できるため、 **営業・マーケティング施策** を集中的に行いやすくなります。 たとえば、 **優先ユーザーには追加クーポンや特別なプランの案内** を行うなど、リソースを最適配分することで収益拡大が期待できます。

5. まとめ

- *欠損補完*や*SMOTE*などを活用し、不均衡データでも精度が出やすい仕組みを整備。
- *introduction*などのテキストも活用して、*依存度*や*経済的余裕*といった潜在的要素を数値化。
- 4種類のモデルを比較し、最適モデルをチューニング後に運用。
- 予測結果を基に *「課金しそうなユーザー」* に優先アプローチを実施。

今後も実際の施策結果をフィードバックしながら、継続的にモデル精度の向上が可能です。