

データ分析課題

Telco Customer Churn

目次

- ▶ 分析概要と目的
- ▶ データ概要
- ▶ 分析タスク01-モデル学習・評価
- ▶ 分析タスク02-施策提示
- ▶ 今後の課題
- ▶ Appendix



01

分析概要と目的

▼ 分析概要と目的

概要と目的

あなたは、電話会社に勤務するデータ分析担当者です。カスタマーサクセス部から契約顧客の解約率の高さが課題になっていると相談を受けており、何かしらの施策を打つ必要があります。

社内で収集したデータを活用し、下記のタスクを実行してください。

タスク実行後は、事業部に対して報告を行ってください。

1. 解約する可能性の高そうな顧客を判別するモデルを作る
2. 解約率を下げるための施策を提示する

データセットはkaggleのTelco Customer Churnを使用する。

02

データ概要

データ概要

【図表1】 データセット(一部)

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No

TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	No	No	One year	No	Mailed check	56.95	1889.5	No
No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
Yes	No	No	One year	No	Bank transfer (automatic)	42.3	1840.75	No
No	No	No	Month-to-month	Yes	Electronic check	70.7	151.65	Yes

使用データ： WA_Fn-UseC_-Telco-Customer-Churn.csv

データ件数：7043件

特徴量(カラム)の数：21カラム

解約者数：1869人(約0.27)

非解約者数：5174人(約0.73)

▼ 特徴量(カラム)概要

【図表2】特徴量概要

特徴量	説明
customerID	顧客固有のID
gender	性別 (男性、女性)
SeniorCitizen	高齢者かどうか (1=はい、 0=いいえ)
Partner	パートナーの有無 (はい、 いいえ)
Dependents	扶養家族の有無 (はい、 いいえ)
tenure	在籍期間(月)
PhoneService	電話サービスの有無 (はい、 いいえ)
MultipleLines	複数回線の有無 (はい、 いいえ、 電話サービスなし)
InternetService	インターネット・サービス・プロバイダー (DSL、光ファイバー、いいえ)
OnlineSecurity	オンライン・セキュリティの有無 (はい、 いいえ、 インターネット サービスなし)
OnlineBackup	オンライン・バックアップの有無 (はい、 いいえ、 インターネット サービスなし)
DeviceProtection	デバイス保護機能の有無 (はい、 いいえ、 インターネット サービスなし)
TechSupport	テクニカル・サポートの有無 (はい、 いいえ、 インターネット サービスなし)
StreamingTV	ストリーミングTVサービスの有無 (はい、 いいえ、 インターネット サービスなし)
StreamingMovies	ストリーミング映画サービスの有無 (はい、 いいえ、 インターネット サービスなし)
Contract	契約形態 (月単位、1年単位、2年単位)
PaperlessBilling	ペーパーレス請求の有無 (はい、 いいえ)
PaymentMethod	お支払い方法 (電子小切手、郵送小切手、銀行振込(自動)、クレジットカード(自動))
MonthlyCharges	月額料金(ドル)
TotalCharges	総支払額(ドル)
Churn	解約の有無 (はい、 いいえ)

03

分析タスク01

- モデル学習・評価

▼ 分析タスク01

分析タスク01

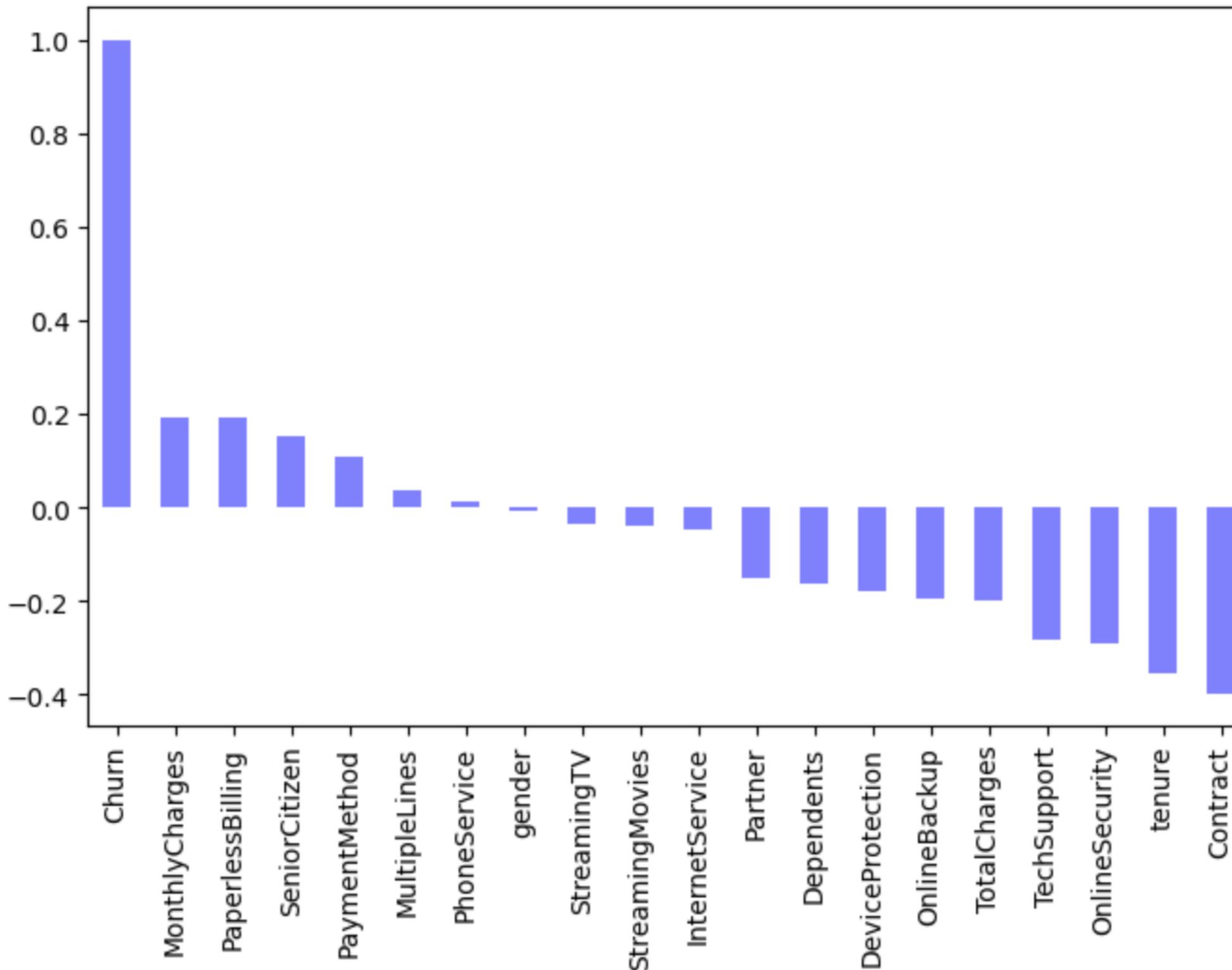
解約する可能性の高そうな顧客を判別するモデルを作る

前提

- ✓ データ分析：データの特性や構造を理解するために、探索的データ分析(EDA)を実施する。
- ✓ モデル：データセットに最も適したアルゴリズムを見つけるため、複数の分類モデル(ロジスティック回帰、サポートベクターマシン、決定木、ランダムフォレスト、XGBoost、LightGBM)を試行し、精度が最も高いモデルを選択する方針とした。
- ✓ 評価指標：モデルの精度を決める評価指標は再現率(Recall)を使用する。解約したユーザーのうち、モデルが正確に特定できたユーザーの数を示す指標である。解約するユーザーを見逃さないようにしたいため、重視すべき指標とした。
- ✓ 目標精度：再現率80%以上を目標とする。再現率80%の場合、解約するユーザーの80%は予測できるモデルということである。

探索的データ分析(EDA)

【図表3】「Churn(解約)」と特徴量との相関

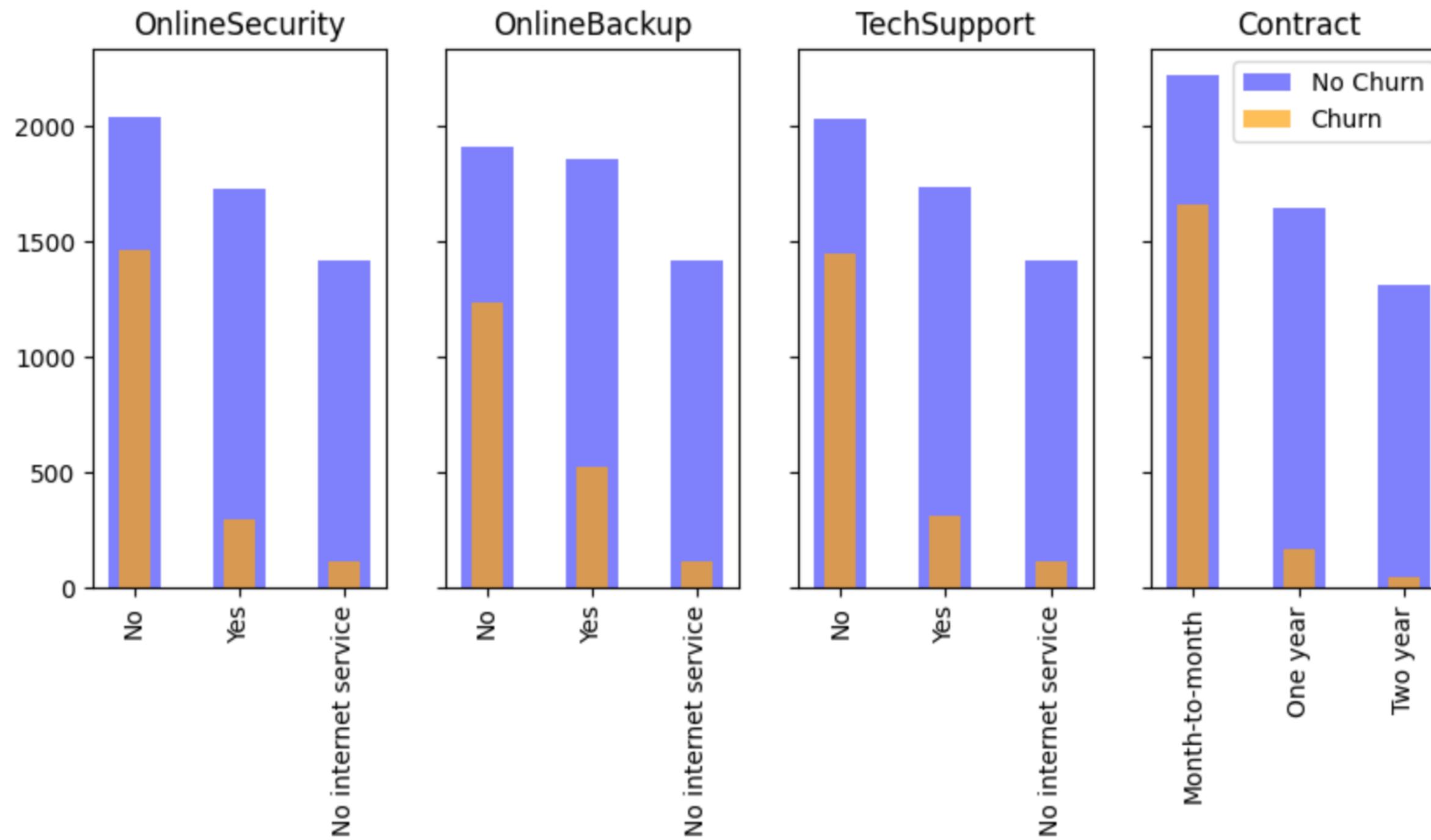


相関分析を実施しChurn(解約)と特徴量との相関を確認した。(図表3)
解約と相関が強かったのは以下の特徴量であった。

- Contract(契約形態)
- tenure(在籍期間)
- OnlineSecurity(オンライン・セキュリティの有無)
- TechSupport(テクニカル・サポートの有無)
- TotalCharges(総支払額)
- OnlineBackup(オンライン・バックアップの有無)
- MonthlyCharges(月額料金)

探索的データ分析(EDA)

【図表4】「Churn(解約)」とカテゴリ特徴量

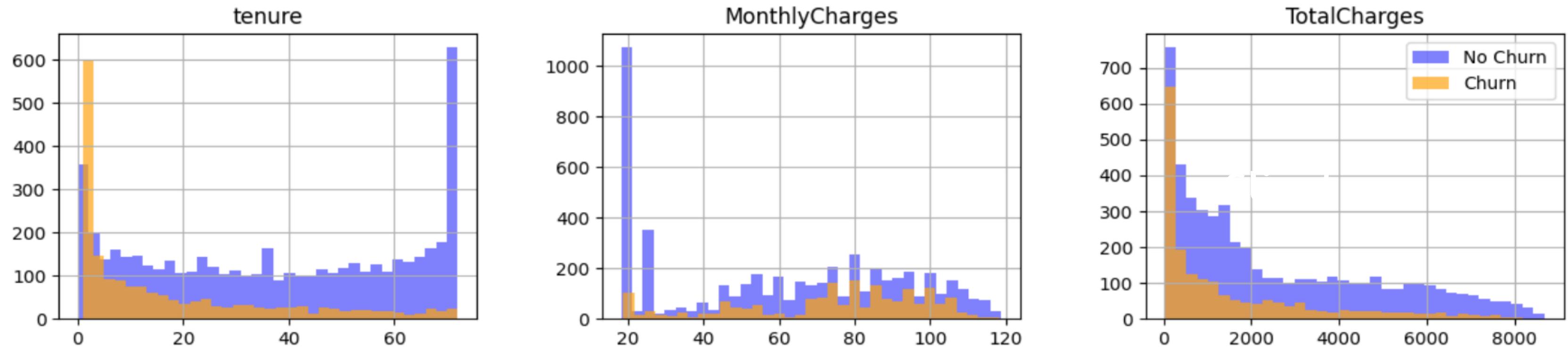


図表3でChurn(解約)と相関が強かった
カテゴリ特徴量をグラフにした。

- インターネット・サービスに加入しているユーザーでは、オンライン・セキュリティやオンライン・バックアップ、テクニカル・サポートに加入していないユーザーは、解約する可能性が高い。
- Contract(契約形態)では月単位のユーザーは、解約する可能性が高い。

探索的データ分析(EDA)

【図表5】「Churn(解約)」と数値特徴量



図表3でChurn(解約)と相関が強かった数値特徴量をグラフにした。

- tenure(在籍期間)では、1ヶ月～2ヶ月の解約者が最も多い。5ヶ月までにユーザーの多くが解約している。
- MonthlyCharges(月額料金)では、70ドル～105ドルのユーザーの多くが解約している。
- TotalCharges(総支払額)では、500ドル以下のユーザーの多くが解約している。

▼ モデル学習・評価

【図表6】モデル比較

Model	Accuracy	Precision	Recall	F1_Score	AUC
ロジスティック回帰	0.74	0.51	0.80	0.62	0.84
XGBoost	0.77	0.56	0.70	0.62	0.84
ランダムフォレスト	0.77	0.55	0.70	0.62	0.84
サポートベクターマシン	0.77	0.55	0.70	0.61	0.81
決定木	0.75	0.52	0.65	0.58	0.78
LightGBM	0.77	0.56	0.65	0.60	0.83

探索的データ分析結果を基にモデルの学習・評価を行った。モデルの学習結果は図表6のようになつた。

・ロジスティック回帰が、再現率(Recall)80%となり最も評価の高いモデルとなつた。目標精度の再現率80%以上は達成できた。

分析タスク01結果：

評価指標である再現率(Recall)が最も高く、目標精度の再現率80%以上に達しているため、解約予測を行うモデルにロジスティック回帰を採用する。

04

分析タスク02 - 施策提示

▼ 分析タスク02

分析タスク02

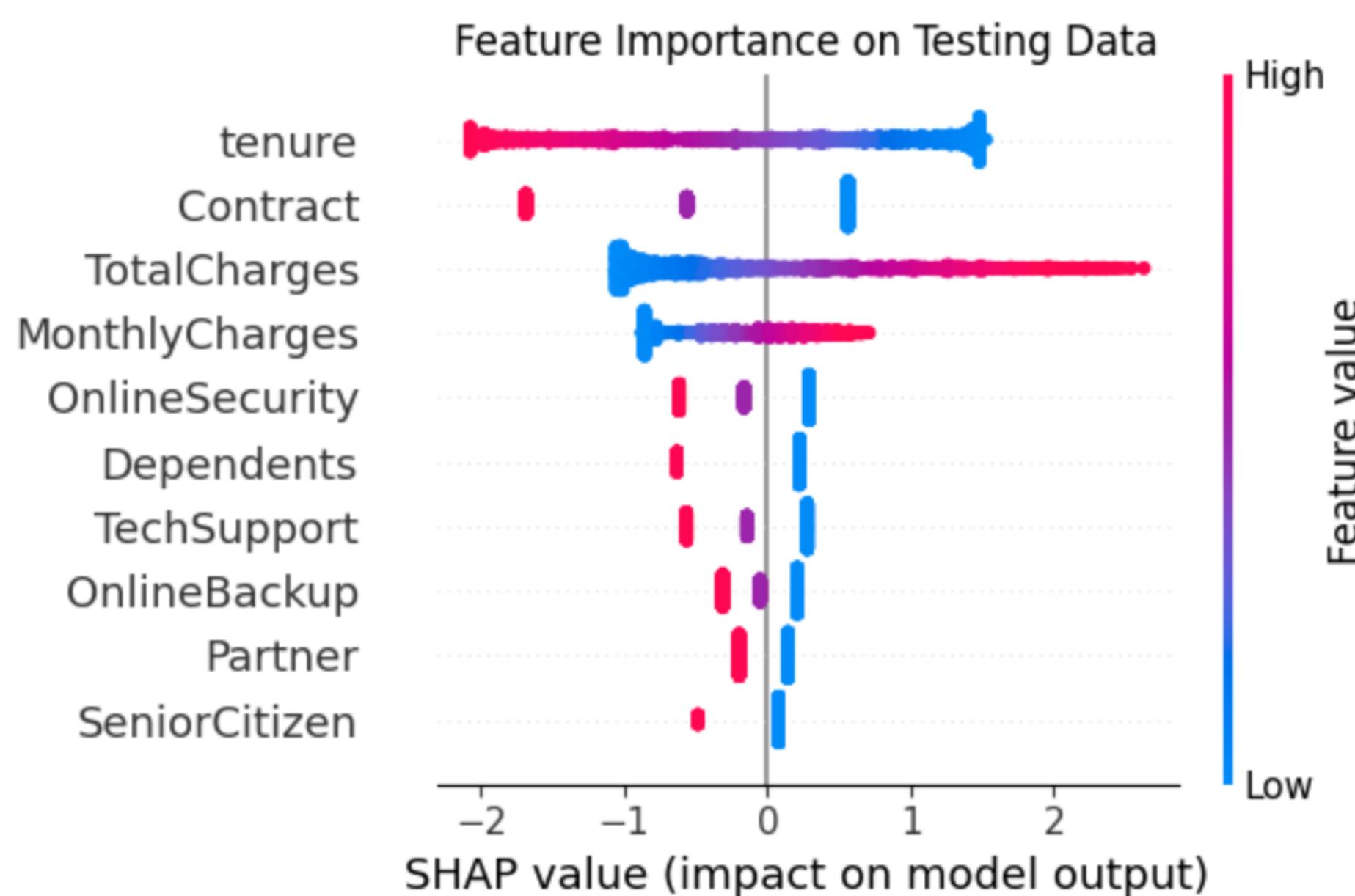
解約率を下げるための施策を提示する

前提

- ✓ 施策は分析タスク01で作成した解約予測モデル(ロジスティック回帰)を使って、解約する可能性が高いと予測したユーザーに対して行う。解約する可能性が高いユーザーに施策を絞れば、費用対効果を高めることができるためである。
- ✓ 分析タスク01で作成した解約予測モデルを分析し、予測に対して影響度の大きい特徴量を抽出する。その特徴量をもとに施策提案を行う。

▼ 各特徴量の寄与度

【図表7】各特徴量の寄与度



各特徴量の解約予測モデルに対する寄与度(上位10個)を可視化した。
(図表7)

- ・図表3の相関分析の結果と近いものになった。図表7の結果をもとに施策を考える。図表3の内容は単純な相関関係であり、精度の高いモデルに影響を与えていたる特徴量の方が重要と考える。モデルが予測したユーザーに対して施策を実施するため、その方が効果的である。

▼各特徴量の寄与度

図表7より影響が大きいと思われる特徴量上位8つを抽出した。

1. tenure(在籍期間)：在籍期間が短いと解約する可能性が高い
2. Contract(契約形態)：月単位の契約だと解約する可能性が高い
3. TotalCharges(総支払額)：総支払額が多いと解約する可能性が高い
4. MonthlyCharges(月額料金)：月額料金が高いと解約する可能性が高い
5. OnlineSecuriy(オンライン・セキュリティの有無)：オンライン・セキュリティに加入していないと解約する可能性が高い(インターネットサービスには加入済み)
6. Dependents(扶養家族の有無)：扶養家族がないと解約する可能性が高い
7. TechSupport(テクニカル・サポートの有無)：テクニカル・サポートに加入していないと解約する可能性が高い(インターネットサービスには加入済み)
8. OnlineBackup(オンライン・バックアップの有無)：オンライン・バックアップに加入していないと解約する可能性が高い(インターネットサービスには加入済み)

▼ インサイト抽出と施策提示

図表7から抽出した上位8つの特徴量より、インサイトを抽出し解約率を下げるための施策を提示する。

1. 在籍期間、契約形態

- ・インサイト：在籍期間が短く契約形態が月単位の契約だと解約率が高い。図表5より、在籍期間5ヶ月までに解約するユーザーが多い。月単位で契約し最初の数ヶ月でサービスに満足しなかったユーザーが解約していると思われる。
- ・**施策①：1年と2年単位の長期契約に割引やクーポンなどの特典を用意する。長期契約者向けに専用サポートラインや優先サービスを提供し、付加価値を高める。特に最初の数ヶ月(5ヶ月)は割引や特典を増やしサービスの質を向上させる。**

2. 総支払額、月額料金、扶養家族の有無

- ・インサイト：総支払額や月額料金が高いユーザーは解約率が高い。料金の負担が解約の一因と考えられる。扶養家族のいないユーザーの多くは、学生や独身者などの若いユーザーと思われる。若いユーザーにとって最も重要なのは料金だと考えられる。図表5より、月額料金が70ドルから105ドルまでのユーザーの解約率が高い傾向にある。
- ・**施策②：料金割引プランや、月額料金の見直しを実施し、学生や若者向けの料金プランの見直しも行う。特に月額料金が70ドルより安い料金プランを増加させる。利用頻度や機能に応じた料金体系も導入する。**

▼ インサイト抽出と施策提示

3. オンライン・セキュリティの有無、テクニカル・サポートの有無、オンライン・バックアップの有無
- ・インサイト：インターネットサービスに加入しているが、上記3つのサービスに加入していないユーザーが解約する可能性が高い。図表4より加入しているユーザーの解約率は低いため、サービスの内容を知らないだけで、加入すれば満足してもらえると思われる。
 - ・**施策③**：広告などサービスを知ってもらうための施策を打つ。パーソナライズされたサービス推奨システムを導入し、各ユーザーの利用パターンに基づいて最適なサービスを提案する。最初の数ヶ月は無料や割引するなど加入特典を設ける。

分析タスク02結果：

分析タスク01で作成した解約予測モデルによって解約すると予測されたユーザーに対して、インサイトが当てはまる場合に、施策①～③を実施することを提案する。
施策の優先順位は影響度の大きさ順の①→②→③とする。

05

今後の課題

今後の課題

分析タスク実施後の課題

1. 解約予測モデルの精度向上

- ・データの増加とともに精度向上の可能性があるため、学習・評価を都度実施していく。
- ・特徴量の種類を増やすことも考える。例えば、「入会時の情報(方法・動機・キャンペーンの有無)」、「サービスの利用頻度」、「外部要因(競合他社のキャンペーンの有無)」などの特徴量があれば予測精度が高まる可能性がある。
- ・アルゴリズムの変更やハイパーパラメータのチューニングなどを実施する。

2. 因果関係の検証

- ・特徴量をもとに抽出したインサイトが正しいのかどうかを確認するために調査(サービス満足度調査や解約者へのインタビューなど)を実施する。Churn(解約)と特徴量の間に因果関係があるかを検証する。

3. 効果検証

- ・実施した施策の効果をABテストなどを行い検証し、今後の分析や施策に適用する。

Appendix

▼ 分析タスク01

データ前処理

1. 数値特徴量への変換

- ・ 数値であるべきTotalCharges(総支払額)がカテゴリ特徴量になっていたため、float型に変換した。変換後、欠損値が出現したため、中央値で補完を行った。

2. customerID(顧客固有のID)の除外

- ・ 個人を識別するただのIDのため、モデル作成には必要ないと判断し除外した。

3. カテゴリ特徴量のエンコーディング

- ・ カテゴリ特徴量はそのままでは機械学習アルゴリズムに適用ができないため、ラベルエンコーディングを実施し、YesやNoなどのカテゴリ特徴量を0や1などの数値に変換した。

▼ 分析タスク01

特徴量エンジニアリングと不均衡データへの対応

1. 特徴量の除外

- Churn(解約)に対する各特徴量の相関分析を行い(図表3)、相関を示さなかった特徴量の Multiplelines(複数回線の有無)、PhoneService(電話サービスの有無)、gender(性別)、StreamingTV(ストリーミングTVサービスの有無)、StreamingMovies(ストリーミング映画サービスの有無)、InternetService(インターネット・サービス・プロバイダー)を除外した。

2. 不均衡データへの対応

- 解約者(約0.27)と非解約者(約0.73)の比率がアンバランスであった。データのバランスを確保するために、SMOTEを使用し均衡データを作成した。評価に影響が出ないようにするため、学習データのみ解約者データを増やし調整した。

3. 特徴量スケーリング

- 数値特徴量に正規化を実施した。異なる単位のデータを小さな共通レンジに落とし込むことで単位の違いの影響を避けるようにした。

▼ 分析タスク01

モデル学習・評価

1. 汎化性能の評価

- ・過学習を防ぐために交差検証を実施した。手法はK-分割交差検証を採用した。データセットを5つに分類し、モデルの学習と評価を5回行い、得られた5回の評価値の平均をとった値を最終的なモデルのスコアとした。より頑健な汎化精度の評価方法である。

2. ハイパーパラメータのチューニング

- ・グリッドサーチを実施した。こちらもモデルの汎化性能を向上させる手法である。指定したパラメータの全ての組み合わせに対して学習を行い、最も良い精度を示したパラメータを採用するものである。

3. 目標精度

- ・評価指標である再現率(Recall)を上げるために、解約判定の閾値を下げる調整した結果、目標精度に達することができた。

▼ 分析タスク02

モデル解釈

1. 各特徴量の寄与度

- SHAPを使用した。SHAPは、既存の解釈手法を統一し、理論的に堅牢で直感的な説明を提供する手法である。モデルに依存せず、どの機械学習モデルにも適用可能で、一貫性のある特徴重要度を計算できるため採用した。

- 学習データとテストデータの両方で特徴量の重要度を確認した。図表7がテストデータを使ったもので、図表8が学習データを使ったものである。数か所特徴量の位置が入れ替わってはいるが、ほぼ同じ結果となった。

【図表8】各特徴量の寄与度

