

Prediction **One**

ユーザーマニュアル

Ver. 1.0

ソニーネットワークコミュニケーションズ株式会社

著作権情報

本書は著作権法で保護された著作物であり、その全部または一部を許可なく複製すること、複製物を配布すること、他のコンピューター用へ変換することおよび他の言語へ翻訳することは著作権の侵害です。

ご注意

予告なく本書の一部または全体を修正、変更することがあります。

商標表示

本書に記載されている会社名、製品名およびサービス名は各社の商標または登録商標です。

更新履歴

Version	更新日	内容
1.0	2019 年 6 月 12 日	初版発行

目次

更新履歴	2
目次	3
はじめに	5
Prediction One とは	5
予測分析とは	6
動作環境	7
ファイル配置	8
チュートリアル	9
データセット	10
データセットと予測分析の各処理の説明	10
データセットの形式について	11
操作の説明	13
プロジェクトとモデルの説明	13
プロジェクトを作成する	13
新規に学習・評価を実行する (モデルを作成する)	13
作成済みのモデルをベースに学習・評価を実行する	14
予測を実行する	14
作成済みモデルの評価結果を閲覧する	14
作成済みモデルの項目の寄与度を閲覧する	14
作成済みモデルの学習設定を確認する	15
コマンドラインから予測を実行する	15
各画面の説明	16
プロジェクト一覧画面	16
設定画面	17

プロジェクト設定画面	19
メイン画面 (概要・共通部分)	20
新規モデル作成画面	22
学習設定画面	23
学習中画面	26
評価結果画面	27
精度評価の詳細画面	29
寄与度の詳細画面	32
学習設定確認画面	33
予測画面	34
予測設定画面	35
予測プレビュー画面	37
予測ファイルの説明	38
コマンドライン機能	41
学習 (learn) コマンド仕様	41
予測 (predict) コマンド仕様	42
項目情報生成(mkcolinfo)コマンド仕様	43
FAQ	45
Appendix: 評価値に関する参考資料	48

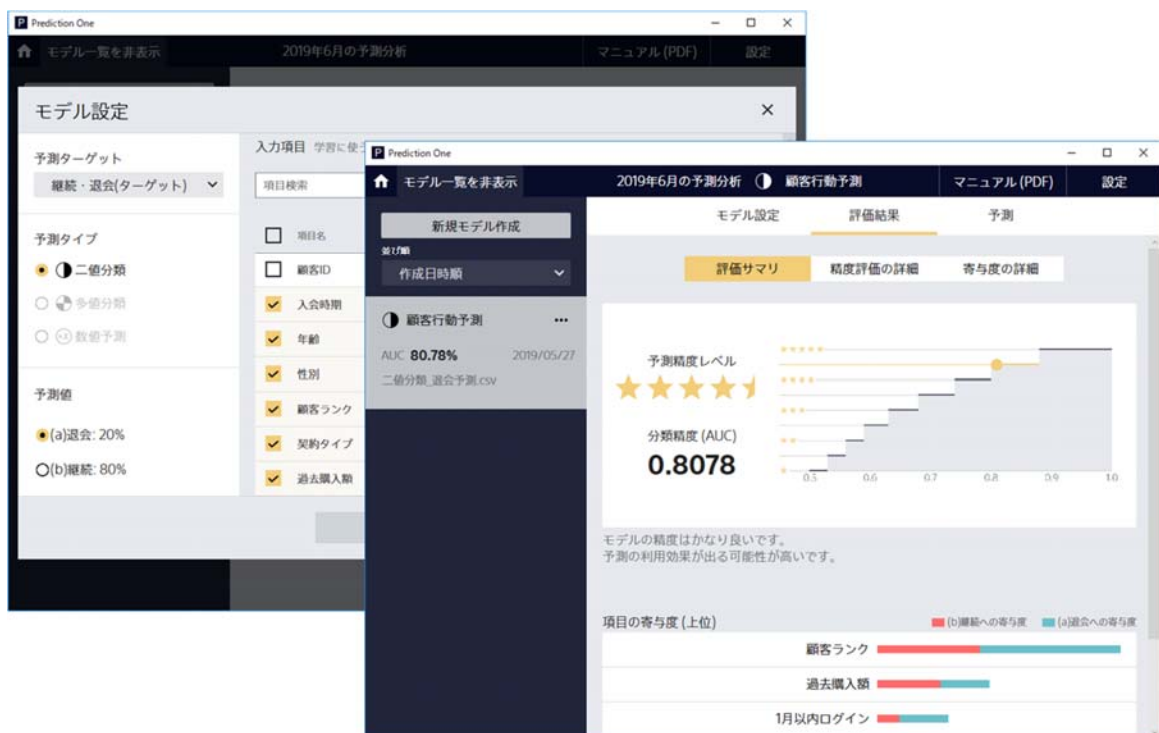
はじめに

Prediction One とは

Prediction One (プレディクションワン) とは、機械学習やプログラミングのスキルなしでも予測分析が実行できる Windows の GUI ソフトウェアです。以下の特長を持ち、使い勝手のよさと、ビジネス導入に必要な予測分析機能を両立しています。

- ✓ **高精度**：データ前処理ノウハウ、自動モデリングアルゴリズムを搭載
- ✓ **使いやすい**：専門用語を排し、利用者のワークフローを考慮したデザイン
- ✓ **理解が容易**：予測寄与度を可視化する機能、豊富な予測精度評価グラフ
- ✓ **導入が容易**：PC にインストールするだけで利用開始、クラウド上へのデータアップロードは不要なので社外秘情報も扱いやすい

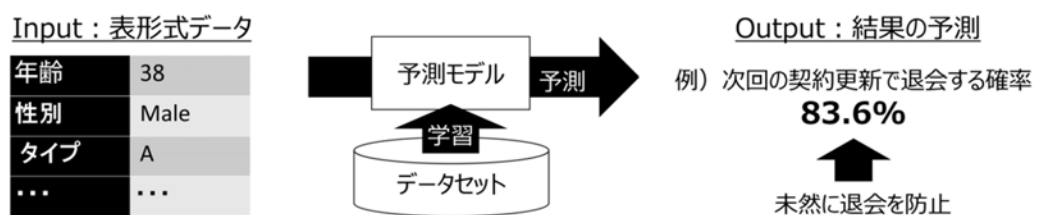
予測分析は汎用性が高く様々なビジネスシーンに導入が可能な技術です。しかし、機械学習などの高い専門性を要するため、適切に導入するにはハードルがあります。Prediction One を利用すれば、機械学習の専門家でなくても、予測分析をビジネスに導入する検討を進められます。



予測分析とは

予測分析とは、過去のデータに基づいて将来の結果を予測する機械学習技術です。例えば、定額制サービスにおいて各顧客の退会確率を事前に予測し、退会確率の高い顧客に対して施策を実施することで退会を減らす、といったことが可能になります。

予測分析とは：過去のデータに基づき将来の結果を予測する機械学習技術



以下のように幅広い応用先が考えられます。

ユースケース	予測	利用方法と効果
ダイレクトマーケティング・CRM	コンバージョン、退会などの顧客行動	予測に基づく顧客セグメントにより ROI を改善
訪問営業	成約などの顧客行動	成約確率の高い顧客を優先的に訪問し効率向上
マッチング、推薦	購入などの顧客行動	購入確率の高い商品を顧客に提示し、売上増加
需要予測	需要の予測	予測を活用し、生産計画の精度を向上
コールセンタ入電予測	入電数予測	入電数予測に基づく、オペレータ配置の精度向上
テキスト分類	苦情・レポートなどの分類	分類の工数削減
人事	業績や退職の予測	従業員の業績改善、退職の防止
リスク推定・査定	デフォルト予測	査定の精度向上、自動化
故障予防	故障の予測	製造機器などの故障の未然防止
不正検知	不正の検知	不正の高精度な予測による不正の削減

予測分析には、以下の予測の種類があります。Prediction One ではすべてに対応しています。

予測タイプ	説明
二値分類	2つの事柄に対する予測確率を算出します。ある事が起きる確率を予測するケースに利用されます。例えば、購入するか購入しないか(購入確率)、故障するか故障しないか(故障確率)、などの予測ができます。
多値分類	3つ以上の事柄に対する予測確率を算出します。例えば、苦情データがA.製品に関するもの、B.サービスに関するもの、C.その他、などの分類の自動化に利用できます。
数値予測 (回帰)	数値を予測します。例えば、来月の出荷台数、来月の入電数などを予測できます。

動作環境

Prediction One の動作環境は以下の通りです。

- Windows 8.1 (64 ビット版)
- Windows 10 (64 ビット版)

OS は最新の Windows Update が適用されている必要があります。OS の言語設定は日本語のみ対応しています。また、以下のソフトウェアをインストールする必要があります。すでにインストール済みの場合はインストール不要です。

- ① Visual Studio 2019 の Microsoft Visual C++ 再頒布可能パッケージ
- ② Microsoft .NET Framework 4.7.2

①については、Prediction One のインストーラーに同封されており、インストールされていない場合は同時にインストールされます。

②については、以下のリンク先からダウンロードおよびインストールが必要な場合があります。

<https://support.microsoft.com/ja-jp/help/4054530/microsoft-net-framework-4-7-2-offline-installer-for-windows>

Prediction One の推奨環境は以下の通りです。

- メモリ： 8GB 以上
- CPU： Core i7
- HDD・SSD： 50GB 以上の空き

ファイル配置

Prediction One の各種ファイル、ワークスペースは以下のフォルダに配置されています。

実行ファイル (GUI)	C:\Program Files\Sony\Prediction One\PredictionOne.exe
実行ファイル (コマンドライン)	C:\Program Files\Sony\Prediction One\PredictionOneCmd.exe
マニュアル	C:\Program Files\Sony\Prediction One\ja-JP\doc\Prediction_One_Document.pdf
チュートリアル	C:\Program Files\Sony\Prediction One\ja-JP\doc\Prediction_One_Tutorial.pdf
サンプルデータセット	C:\Program Files\Sony\Prediction One\ja-JP\doc\sample_dataset\ <ul style="list-style-type: none">● 数値予測_需要予測.csv● 数値予測_需要予測（予測用）.csv● 多値分類_苦情分類.csv● 多値分類_苦情分類（予測用）.csv● 二値分類_故障予測.csv● 二値分類_故障予測（予測用）.csv● 二値分類_退会予測.csv● 二値分類_退会予測（予測用）.csv
ワークスペース	C:\Users\[ユーザー名]\AppData\Local\Sony Corporation\Prediction One\WorkSpace

チュートリアル

同封のチュートリアルファイルをご参照ください。使用を始める前にご覧になることをお勧めします。

C:\Program Files\Sony\Prediction One\ja-JP\doc\Prediction_One_Tutorial.pdf

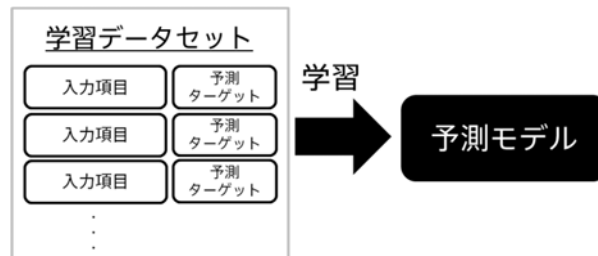
GUI ソフトウェアの設定画面からもファイルを開くことができます。

データセット

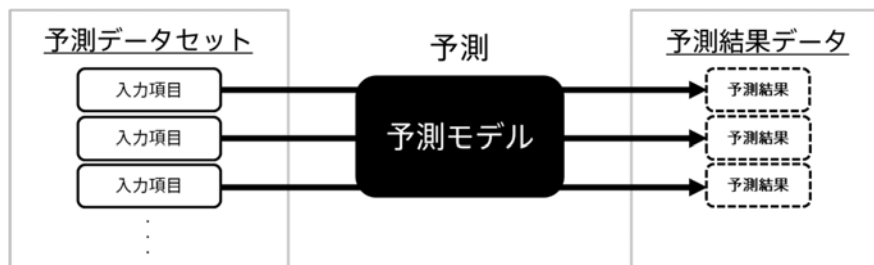
データセットと予測分析の各処理の説明

データセットとは、予測分析の学習、予測、評価に用いるデータの集まりです。Prediction One を利用する前に、準備する必要があります。

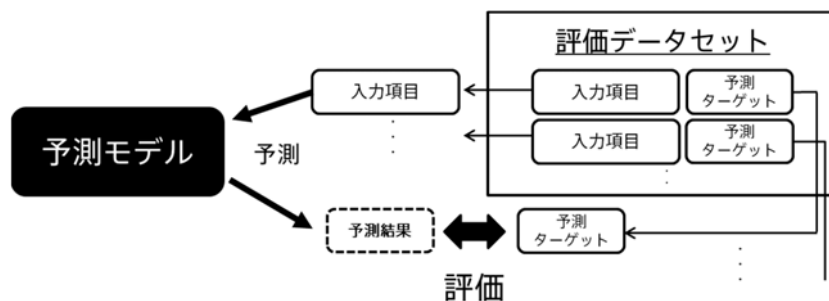
学習とは、学習データセットを与えて、入力項目に対して予測ターゲット項目の値を予測するモデル (関数) を作成することです。



予測とは、学習済みのモデルを利用して、入力項目から予測結果を算出することです。



評価とは、評価データセット (一般に学習とは別のデータセットを利用する) を利用して、予測結果と予測ターゲット項目の実際の値を比較することで学習済みのモデルの性能を評価することです。



データセットの形式について

データ形式としては表形式データ、ファイルフォーマットとしてはCSV(カンマ区切り)とTSV(タブ区切り)に対応しています。

各行が1つのインスタンス(例えば、顧客)に対応し、各列(項目)がインスタンスの属性(例えば、年齢、性別など)に対応します。データファイルの最初の行は列名(項目名)を表す内容を書き、2行目以降にインスタンスの情報を書きます。**各行は同じ項目数である必要があります。**

予測ターゲットの項目(例えば、継続か退会か)は項目の1つとして、ファイルに書かれているとします。データセットの例として、下図をご参照ください。予測ターゲット項目は「継続・退会」です。同封のサンプルデータも例として参考にしてください。

顧客ID	年齢	性別	顧客セグメント	契約タイプ	過去購入額	登録日	1月以内ログイン	継続・退会
id1	28		A	契約A		2018/3/4	あり	(b)継続
id2	43	男性	C	契約B	8000	2017/8/11	なし	(a)退会
id3	44		A	契約B		2017/1/12	あり	(b)継続
id4	33	女性	A	契約A	4000	2016/5/4	なし	(b)継続
id5	36	男性	B	契約B	8000	2018/9/10	あり	(b)継続
id6	38	男性	D	契約A	4000	2016/2/4	あり	(b)継続
id7	36	男性	B	契約B		2017/3/3	あり	(b)継続

項目としては以下のデータタイプを利用できます。

タイプ	説明
文字列	カテゴリカル値(例：上記の「性別」項目)やテキスト(日本語と英語に対応)
数値	整数や小数などの数値(例：上記の「過去購入額」項目)
日付	日時などの日付(例：上記の「登録日」項目)

欠損値がある場合は、空文字を利用してください。

Prediction One では、ファイルの先頭 1000 行を読み込んで予測ターゲット項目の値が何種類あるのかを判定します(上記の例だと「継続」と「退会」の2種類)。予測ターゲット項目が文字列の場合、判定されたユニーク数によって2値分類か多値分類かの判定を行います。

二値分類を実行したい場合、先頭 1000 行に予測ターゲット項目の値が2種類出現するようにしてください。多値分類を実行したい場合、先頭 1000 行に予測ターゲット項目の値が3種類以上出

現するように並び替えておいてください。また、20 種類を超える分類には対応していませんのでご注意ください。

操作の説明

本章の内容は、次章の「各画面の説明」を参照しながら読んでください。

プロジェクトとモデルの説明

プロジェクトとは

Prediction One では、ユーザーが自分の作業単位に合わせて、プロジェクトを作成します。1つのプロジェクト内で複数のモデルを管理することができます。

モデル (予測モデル) とは

Prediction One では、学習と評価が同時に実行された後、予測モデルが作成されます。作成された予測モデルは関連情報 (精度評価結果など) と合わせて保存されます。作成されたモデルに対して、予測データを入力することで、予測結果を算出できます。

プロジェクトを作成する

1. プロジェクト一覧画面に遷移します。
2. 「新規プロジェクト」ボタンをクリックします。
3. プロジェクト設定画面で、プロジェクト名等を入力し、「決定」ボタンを押します。

新規に学習・評価を実行する (モデルを作成する)

1. プロジェクトのメイン画面に遷移します。
2. 「新規モデル作成」をクリックします。
3. データ読み込み画面で、学習データセットが書かれたファイルを指定し、読み込みます。
 - CSV・TSV 以外のファイルは指定しないでください。
4. 学習設定画面で予測ターゲットを指定します (必要に応じて予測タイプ等も指定)。
5. 「学習と評価を実行」ボタンをクリックします。

作成済みのモデルをベースに学習・評価を実行する

1. プロジェクトのメイン画面に遷移します。
2. 「モデル一覧」から、ベースにしたいモデルのプルダウンボタンをクリックします。
3. 「コピーして新規モデルを作成」をクリックします。
 - ベースのモデルを作成した際に利用したデータが、同じ場所に同じファイル名で配置されている場合のみ可能な操作です。ベースのモデル作成後にファイル内容を変更していた場合、その後の処理が正しく動作しない場合があります。
4. 学習設定画面で設定を変更します。
5. 「学習・評価を実行」ボタンをクリックします。

予測を実行する

1. プロジェクトのメイン画面に遷移します。
2. 「モデル一覧」から、予測に使用したいモデルを選択します。
3. 「予測」タブをクリックします。
4. 「新規予測」ボタンをクリックします。
5. データ読み込み画面で予測データセットが書かれたファイルを指定します。
6. 必要に応じて、オプションをチェックし、「決定」ボタンを押します。
7. 予測プレビュー画面で、「予測結果を保存」をクリックし、予測結果ファイルを保存します。

作成済みモデルの評価結果を閲覧する

1. プロジェクトのメイン画面に遷移します。
2. 「モデル一覧」から、評価結果を閲覧したいモデルを選択します。
3. 「評価結果」タブをクリックします。
4. 「評価サマリ」タブや「精度評価の詳細」タブをクリックします。

作成済みモデルの項目の寄与度を閲覧する

1. プロジェクトのメイン画面に遷移します。
2. 「モデル一覧」から、寄与度を閲覧したいモデルを選択します。

3. 「評価結果」タブをクリックします。
4. 「寄与度の詳細」タブをクリックします。

作成済みモデルの学習設定を確認する

1. プロジェクトのメイン画面に遷移します。
2. 「モデル一覧」から、学習設定を閲覧したいモデルを選択します。
3. 「モデル設定」タブをクリックします。

コマンドラインから予測を実行する

1. プロジェクトのメイン画面に遷移します。
2. 「モデル一覧」から、予測に使用したいモデルを選択します。
3. 「予測」タブをクリックします。
4. 「新規予測」ボタンの下の文から、選択中のモデルのフォルダ名を確認します。
5. 本マニュアルの「コマンドライン機能」を参考にコマンドを作成します (上記のモデルフォルダ名を利用します)。
6. コマンドを実行します (コマンドプロンプトなどから実行します)。

各画面の説明

プロジェクト一覧画面

Prediction One を起動すると表示される画面です。この画面ではプロジェクトの一覧が表示されます。他の画面では左上の Home ボタンをクリックすると、この画面に戻ることができます。



プロジェクトを作成するには

新規プロジェクトボタンをクリックします。新規プロジェクト設定画面に遷移しますので、そこで必要な情報を入力します。

プロジェクトを開くには

プロジェクトタイルをクリックします。

プロジェクト名やプロジェクトの説明を変更するには

変更したいプロジェクトのプロジェクトプルダウンから編集を選択します。プロジェクト設定画面に遷移しますので、そこでプロジェクト名を変更します。

プロジェクトを削除するには

削除したいプロジェクトのプロジェクトプルダウンから削除を選択します。一度削除すると復元できませんので、ご注意ください。

設定画面を開くには

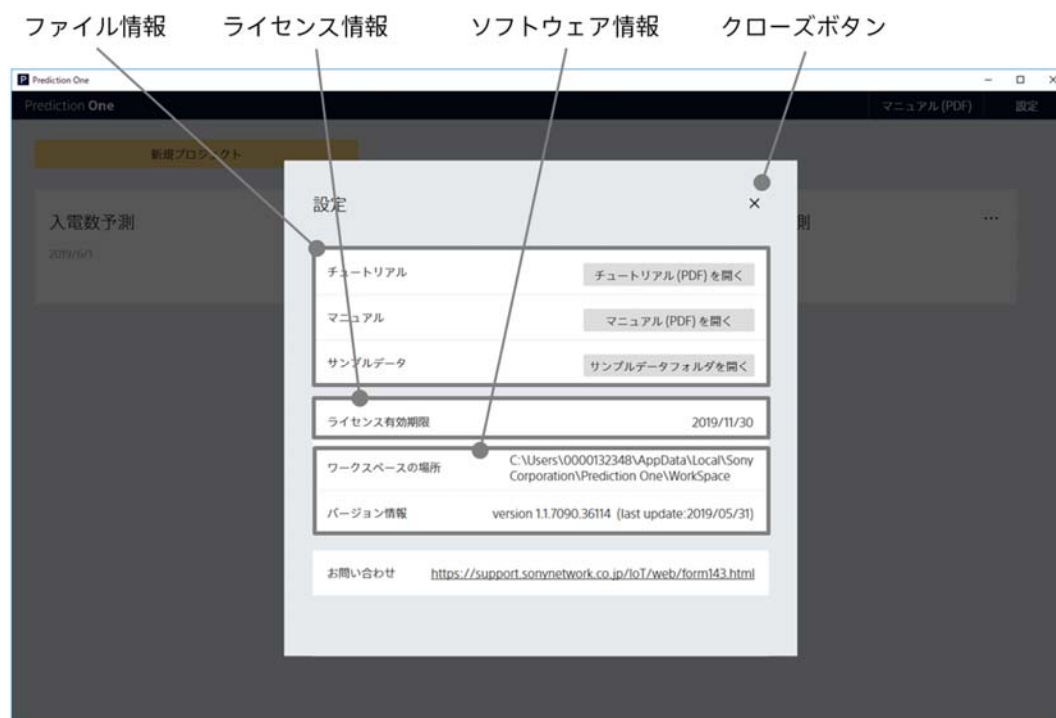
設定ボタンをクリックします。他の画面でも同様です。詳細は「設定画面」の説明をご覧ください。

マニュアルを開くには

マニュアルボタンをクリックします。他の画面でも同様です。

設定画面

各画面において、右上の設定ボタンを押すところの画面に遷移します。



チュートリアルを開くには

ファイル情報の「チュートリアル (PDF) を開く」ボタンをクリックしてください。

マニュアルを開くは

ファイル情報の「マニュアル (PDF) を開く」ボタンをクリックしてください。

サンプルデータが格納されているフォルダを開くには

ファイル情報の「サンプルデータフォルダを開く」ボタンをクリックしてください。

ライセンスの有効期限を確認するには

ライセンス情報に記載の「ライセンス有効期限」を確認してください。

ワークスペースの場所を確認するには

ソフトウェア情報の「ワークスペースの場所」を確認してください。ワークスペースとは、Prediction One のプロジェクトやモデルなどの情報が格納されているフォルダです。

バージョン情報を確認するには

ソフトウェア情報の「バージョン情報」を確認してください。

設定画面を閉じるには

クローズボタンをクリックしてください。

プロジェクト設定画面

プロジェクト一覧画面で新規プロジェクトボタンをクリックするとこの画面に遷移します。プロジェクトプルダウンから編集を選択した際も同様の画面に遷移し、プロジェクト名や説明の編集ができます。



プロジェクト名を入力するには

プロジェクト名テキストボックスにテキストを入力し、決定ボタンを押します。プロジェクト名には改行を入力できません。

プロジェクトの説明を入力するには

プロジェクトの説明をテキストで追加したい場合は、プロジェクト説明テキストボックスに入力してください。

プロジェクト一覧画面に戻るには

キャンセルボタンまたはクローズボタンをクリックしてください。

プロジェクトを作成するには

プロジェクト名を入力し、決定ボタンをクリックしてください。

メイン画面 (概要・共通部分)

この画面はプロジェクト一覧画面でプロジェクトをクリックすると表示されます。または新規プロジェクト作成後に表示されます。

メイン画面では、主に

- モデルを作成する (学習する) こと
- 作成済モデルの設定と精度評価結果、寄与度分析結果を閲覧すること
- 作成済モデルに対して予測処理を実行すること

ができます。

Homeボタン モデル一覧表示切替ボタン 新規モデル作成ボタン モデル設定タブ 評価結果タブ 予測タブ



プロジェクト一覧に戻るには

Home ボタンをクリックしてください。

新規にモデルを作成するには

新規モデル作成ボタンをクリックしてください。

モデル一覧の表示・非表示を切替えるには

モデル一覧表示切替ボタンをクリックしてください。

過去に作成したモデルを閲覧するには

モデル一覧をご覧ください。詳細を見たいモデルのモデルタイルをクリックしてください。クリックしたモデルに関する内容に切り替わります。

モデルリストをソートするには

モデルソートプルダウンからソート基準の指標を選択してください。デフォルトでは、作成日時順に並んでいます。

モデルのモデル名や説明を編集するには

モデルプルダウンをクリックし、表示されたメニューから「編集」をクリックしてください。

モデルを削除するには

モデルプルダウンをクリックし、表示されたメニューから「削除」をクリックしてください。

作成済みのモデルを編集し、再度学習を実行するには

元となるモデルのモデルプルダウンをクリックし、「コピーして新規モデルを作成」をクリックしてください。元となるモデルの設定が入力された状態で、学習設定画面を開くことができます。ただし、元となるモデルの作成時に利用した学習データが同じ場所に同じファイル名で存在しない場合、この操作は実行できません。また、元となるモデル作成後に学習データのファイル内容を変更していた場合、その後の処理が正しく動作しない場合があります。

モデルの設定を確認するには

モデル設定タブをクリックしてください。詳細は「学習設定確認画面」をご覧ください。

モデルの精度評価結果や寄与度分析結果を確認するには

評価結果タブをクリックしてください。詳細は「評価結果画面」をご覧ください。

予測を行うには

予測タブをクリックしてください。詳細は「予測画面」をご覧ください。

設定画面を開くには

設定ボタンをクリックします。詳細は「設定画面」をご覧ください。

マニュアルを開くには

マニュアルボタンをクリックします。

新規モデル作成画面

新規モデル作成ボタンをクリックするとこの画面に遷移します。



学習データを指定 (ファイルを指定する) には

中央のウィンドウの領域にファイルをドラッグすることで指定することができます。別の方法として、ファイル指定ボタンをクリックし、ファイルブラウザから指定することもできます。ファイルフォーマットは CSV (カンマ区切り) と TSV (タブ区切り) に対応しています。

モデル名を変更するには

モデル名はデフォルトの名称が自動的に入力されます。モデル名テキストボックスにテキストを入力・変更し、決定ボタンを押します。改行は入力できません。最大文字数の制限もあります。

モデルの説明を入力するには

モデル説明テキストボックスに入力してください。

前の画面に戻るには

キャンセルボタンまたはクローズボタンを押してください。

学習設定画面

この画面では、学習データ読み込み結果が表示され、学習の設定を入力します。「新規モデル作成画面」でのデータ読み込み後、またはモデルタイルで「コピーして新規モデルを作成」クリック後に遷移します。

予測ターゲットプルダウン 予測タイプ選択 予測値選択 項目検索ボックス 項目リスト

項目名	データタイプ	ユニーク数 (先頭1000行中)	欠損
<input type="checkbox"/> 顧客ID	文字列	1,000	-
<input checked="" type="checkbox"/> 入会時期	日付	807	-
<input checked="" type="checkbox"/> 年齢	数値	36	-
<input checked="" type="checkbox"/> 性別	文字列	3	-
<input checked="" type="checkbox"/> 顧客ランク	文字列	6	0%
<input checked="" type="checkbox"/> 契約タイプ	文字列	4	-
<input checked="" type="checkbox"/> 過去購入額	数値	257	-
<input checked="" type="checkbox"/> 継続回数	数値	10	3%
<input checked="" type="checkbox"/> メアド登録	文字列	2	-
<input checked="" type="checkbox"/> 1月以内ログイン	文字列	2	-

特徴タイプ	データタイプ	値	割合
契約タイプ	文字列	4	-
過去購入期	数値	257	-
契約回数	数値	10	3%
メイド登録	文字列	2	-
1月以内ログイン	文字列	2	-
契約担当者名	文字列	813	-
契約 - 退会(ターゲット)	文字列	2	-

評価方法選択 交差検証チェック 学習データ変更ボタン キャンセルボタン 学習と評価を実行ボタン

学習・評価を実行するには

予測ターゲットを指定後、「学習と評価を実行」ボタンをクリックしてください。

予測ターゲットを指定するには

予測ターゲットプルダウンから、予測ターゲットとなる項目を選択します。

予測タイプを指定するには

予測タイプ指定から、予測タイプを指定します。一部を除き、以下のルールで自動的に予測タイプが選択されます。ルールは値のユニーク数 (取り得る文字列の種類の数) を利用するのですが、ユニーク数は学習ファイルの先頭 1000 行を読み込んでカウントしています。

- 予測ターゲットのデータタイプが文字列の場合
 - 項目内の値のユニーク数が 2 の場合は二値分類となります。
 - ユニーク数が 3 以上 20 以下の場合、多値分類となります。
 - ユニーク数が 21 以上の場合、Prediction One は多値分類に対応していません。データを変更・加工するなどし、20 以下に変更してください。
- 予測ターゲット項目のデータタイプが数値の場合
 - 項目内の値のユニーク数が 2 の場合は二値分類となります。
 - ユニーク数が 3 以上 20 以下の場合、多値分類と数値予測が選択できます。
 - ユニーク数が 21 以上の場合、数値予測となります。

- **二値分類が選択され、1000 行目より後に予測ターゲット項目に別の値が含まれる場合、学習が途中で失敗して終了します。多値分類が選択され、1000 行目より後のデータも含めた処理時に予測ターゲット項目のユニーク数が 20 を超えると、学習が失敗して途中で終了します。**

モデルの入力として使用する項目を変更するには

項目リストの項目名の左のチェックボックスでチェックされている項目をモデルの入力として利用します。変更したい場合は、学習と評価を実行する前にチェックを変更してください。

ユニーク数が多すぎる項目は学習に悪影響を与える (過学習が起きて予測精度が悪くなる) ことが多いので、デフォルトでチェックを外しています。

項目数が多すぎる場合は、項目検索ボックスに文字を入力すると、項目リストに表示される項目を絞ることができます。

二値分類において、予測値を指定するには

予測タイプとして二値分類を選択すると、予測したい値を選択・変更できます。予測したい方をチェックしてください。学習後の評価において、チェックされた方を中心に評価グラフ等が生成されます。

学習データを変更するには

学習データ変更ボタンをクリックし、その後表示される画面から、ファイルを指定してください。

評価データを指定するには

評価データの指定はオプションです。あらかじめ用意したデータで評価を行いたい場合に、「評価方法選択」から「評価データを指定」を選択し、「ファイルを指定」ボタンをクリックし、その後表示される画面でファイルを指定してください。この場合、学習データはすべて学習に利用され、指定された評価データを評価に利用するようになります。**評価データと学習データは項目がすべて一致している必要があります。一致していない場合、学習・評価の処理が失敗します。**

評価方法として交差検証を指定するには

評価データが指定されておらず、データ数が 500 以下の場合、交差検証が実行されます。Prediction One ではデータを 2 分割し、それぞれを学習データ、評価データとして学習と評価を 2 パターン行い、評価値を算出します。

評価データが指定されておらず、データ数が 500 より大きい場合、データを 9 : 1 に分割し、9 割のデータを学習に利用し、1 割のデータを評価に利用します。交差検証チェックをチェックすると交差検証を実行します。

項目リストについて

- データタイプ： 自動判定された項目のタイプが表示されます。タイプは、文字列、数値、日付の 3 種類があります。自動判定には、学習ファイルの先頭 200 行を利用しています。プルダウンの操作によって、数値と日付は、文字列のデータタイプに変更することができます。
- ユニーク数： 項目の取り得る値の種類の数です。学習ファイルの先頭 1000 行を読み込んでカウントしています。
- 欠損： 項目内の欠損の割合を表示しています。空文字列を欠損として扱います。欠損率は、学習ファイルの先頭 1000 行を読み込んでカウントしています。

学習中画面

学習設定画面で「学習と評価を実行」ボタンをクリックすると、この画面に遷移します。

学習・評価フェーズ 学習・評価メッセージ 残り時間 キャンセルボタン (OKボタン)



学習と評価の進捗状況を確認するには

学習・評価時に現在どの処理を実行しているかが学習・評価フェーズに表示されます。

- データ読み込み：ファイルからデータを読み込み、学習用データセットと評価用データセットを準備します。処理時間の見積もりも行います。
- データ前処理：データセットについて必要な統計量を計算します。
- モデル学習：学習データセットから予測モデルを学習します。Prediction One では、モデルとしてニューラルネットワークと勾配ブースティング木を利用しており、それぞれ最適なモデル設定を自動的に選択します。
- 精度評価：評価用データセットを使って、作成した予測モデルの予測精度を計算します。
- 予測寄与度分析：作成した予測モデルの予測寄与度を分析します。

学習・評価メッセージに処理内容の詳細が表示されます。

残り時間を確認するには

下部に見積もった残り時間が表示されます。見積もりは外れる可能性もあります。

学習をキャンセルするには

キャンセルボタンをクリックします。モデル学習のみ処理途中のキャンセルが可能なため、他の処理中の場合は待ち時間が発生する可能性があります。

学習と評価を完了するには

学習と評価の処理が完了後、OK ボタンをクリックしてください。

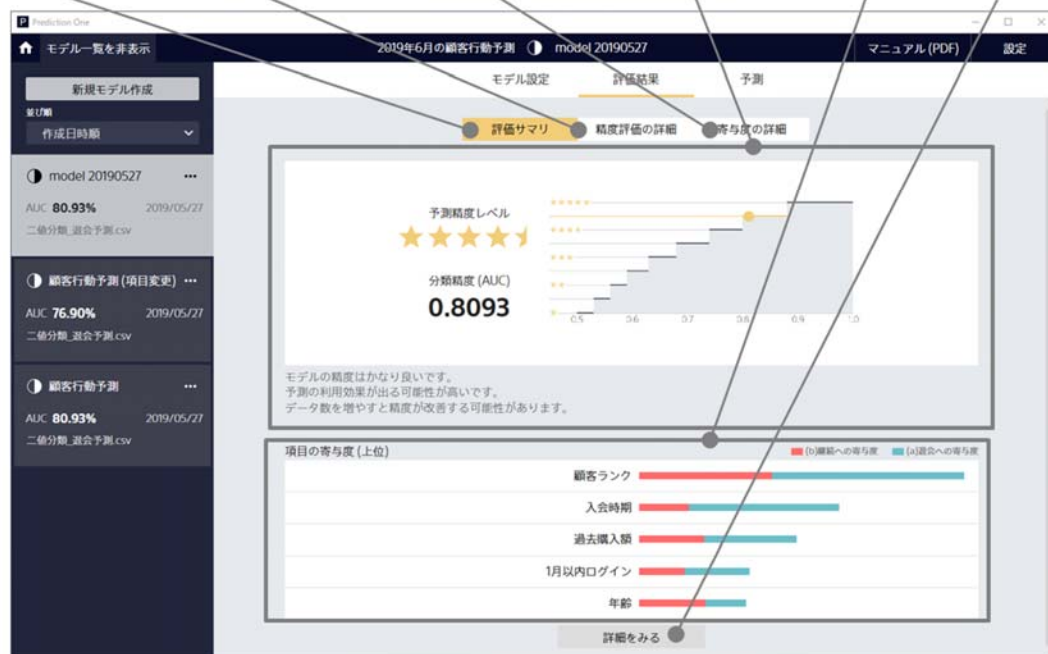
評価結果画面

メイン画面で評価結果タブをクリックするとこの画面に遷移します。

この画面では、学習されたモデルの予測精度などの評価結果を閲覧することができます。基本的に評価では、学習に使用していないデータに対して、学習したモデルで予測を実行し、予測結果と正解の比較により、予測精度を評価します。各予測タイプ (二値分類、多値分類、数値予測) に即した内容が表示されます。

評価サマリタブ、精度評価の詳細タブ、寄与度の詳細タブをクリックすることで表示内容を切り替えることができます。

評価サマリタブ 精度評価の詳細タブ 寄与度の詳細タブ 予測精度サマリ 寄与度サマリ 詳細をみるボタン



予測精度の概要を理解するには

予測精度サマリをご覧ください。各予測タイプの主要な精度評価指標と、予測精度レベル、その説明文が表示されます。予測精度レベルは星の数が多いほど良いことを示します。

データ数を増やすと精度が上がると判定された場合は、その旨がテキストとして表示されます。

寄与度の概要を理解するには

寄与度サマリをご覧ください。各入力項目について、学習した予測モデルがどの程度重要視しているか・予測に有効かを表しています。棒グラフが長いほど重要度・有効度が高いことを表します。重要度・有効度の高い項目が予測結果に大きな影響を与えています。2種類の色は、二値分類の場合は、各値に対する重要度・有効度を表します。数値予測の場合は、予測値を上げる方・下げる方への寄与を表します。

寄与度の詳細を閲覧するには

詳細をみるボタンをクリックしてください。

精度評価の詳細画面

メイン画面で評価結果タブをクリックし、精度評価の詳細タブをクリックするとこの画面に遷移します。



予測精度の評価値の説明を閲覧するには

予測精度の評価値の名称にカーソルを合わせると、ツールチップが表示され、説明を読むことができます。

予測精度のグラフ・表を閲覧するには

評価値の下領域に予測精度を表現するグラフや表が表示されます。スクロールしてご覧ください。

評価値について

予測タイプ	評価値	説明
二値分類	AUC	Area Under the ROC Curve のこと。予測確率のソート順に対する評価。大きいほど精度が良いことを表し、最大 1.0、最小 0.0 である。予測がランダムの場合は 0.5 程度となる。
	Accuracy	大きいほど良い。予測された値と、実際の予測ターゲット項目の値が一致した割合。Appendix も参照のこと。
	Precision	大きいほど良い。予測を、指定した予測値の検出と見た場合、どの程度正確に検出できたかを示す値。大きいほど検出が正確であることを示す。Appendix も参照のこと。
	Recall	大きいほど良い。予測を、指定した予測値の検出と見た場合、検出でどの程度カバーできたかを示す値。大きいほど広くカバーできたことを示す。Appendix も参照のこと。
	F-measure	大きいほど良い。Precision と Recall の総合力を示す値。Appendix も参照のこと。
多値分類	平均 Precision	各予測値 (多値なので 3 つ以上) について、Precision を算出し、平均をとった値
	平均 Recall	各予測値 (多値なので 3 つ以上) について、Recall を算出し、平均をとった値
	平均 F-measure	各予測値 (多値なので 3 つ以上) について、F-measure を算出し、平均をとった値
	Accuracy	大きいほど良い。予測された値と、実際の予測ターゲットの値が一致した割合。Appendix も参照のこと。
数値予測	誤差中央値	小さいほど良い。予測値と実際の値の差の絶対値について、評価データにおける中央値をとった値。Appendix も参照のこと。

誤差率中央値	小さいほど良い。予測値と実際の値の誤差率の絶対値について、評価データにおける中央値をとった値。予測ターゲット項目に絶対値の大きい値が多い場合、予測精度が低くても小さい値になる場合がある。Appendix も参照のこと。
RMSE	小さいほど良い。Root Mean Squared Error のこと。Appendix も参照のこと。
決定係数	大きいほど良い。入力項目が予測ターゲット項目のどれくらいを説明できるか表す値である。Appendix も参照のこと。

混同行列

以下の 4 パターンについて (二値分類かつ取り得る値が A と B の場合)、評価データにおける件数を表示したもの。Appendix も参照してください。

- 予測値が A、実際の値が A ⇒ 正解
- 予測値が A、実際の値が B ⇒ 不正解
- 予測値が B、実際の値が A ⇒ 不正解
- 予測値が B、実際の値が B ⇒ 正解

多値分類はこれらのパターンが多値に拡張されます。

データ分布

二値分類の場合：

予測確率が横軸であり、同じ予測確率を算出されたデータにおける、実際の予測ターゲットの値のデータ数を表示しています。例えば、ターゲットが値 A である場合、予測確率が高いケースで値 A のデータ数が多く、予測値 B (A でない方) のデータ数が少ないと、精度良く予測ができてことになります。

数値予測の場合：

評価対象のデータについて、予測値 (横軸) と実際の値 (縦軸) において、それぞれの (予測値、実際の値) の組プロットしたもの。対角線上に点が並ぶほど良い予測精度であると言えます。

寄与度の詳細画面

メイン画面で評価結果タブをクリックし、寄与度の詳細タブをクリックするとこの画面に遷移します。

学習したモデルが、予測の際にどの入力項目をどの程度どのように参照しているかのサマリを表示します。

- 予測ターゲットの項目にはどの入力項目の関連が強いのか
- 関連が強い項目について、どの値が予測確率を上げるのか下げるのか

といったことを知ることができます。これらは予測モデルの傾向を理解することにも有効ですし、他者に分析結果を説明する際にも利用できます。



項目の寄与度

各入力項目について、予測モデルがどの程度重要視しているか・予測に寄与するかを表しています。棒グラフが長いほど重要度・寄与度が高いことを表します。重要度・有効度の高い項目が予測結果に大きな影響を与えています。2種類の色は、二値分類の場合は、各値に対する重要度・有効度を表します。数値予測の場合は、予測値を上げる方・下げる方への寄与を表します。

項目内容の寄与度

項目の値が、予測ターゲットのどちらにどの程度寄与しているかを、上位3つまで表示しています。項目の寄与度において、項目名を選択することで、対象項目を切り替えることができます。

- 項目内容： 項目の値に相当します。文字列タイプの項目であれば文字列が、数値タイプの項目であれば数値範囲（“[下限値]～[上限値]”の形式）が、日付タイプであれば、年/月/日のいずれかが、表示されます。文字列タイプの項目がテキストとして処理された場合は、単語の原型が表示されます。
- 寄与度 (二値分類)： 予測ターゲットの2種類の値について、予測確率増加に寄与する割合を表しています。
- 寄与度 (数値予測)： 予測値について、増加および減少に寄与する割合を表しています。
- 項目に占める割合：この値が項目に占める割合を表します。この値を持つデータ数の全データ数に占める割合です。

学習設定確認画面

メイン画面でモデル設定タブをクリックするとこの画面に遷移します。

学習設定情報 項目情報

2019年6月の顧客行動予測 顧客行動予測 マニュアル (PDF) 設定

新規モデル作成 作成日時順

顧客行動予測 (項目変更) ...
AUC 76.90% 2019/05/27
二値分類_退会予測.csv

顧客行動予測 ...
AUC 80.93% 2019/05/27
二値分類_退会予測.csv

学習データ
二値分類_退会予測.csv
18153バイト、12列

評価データ
学習データから抽出

項目検索

項目名	データタイプ	ユニーク数 (1670行中)	欠損
顧客ID	数値		
入会時期	日付	1,172	-
年齢	数値	36	-
性別	文字列	3	-
顧客ランク	文字列	6	0%
契約タイプ	文字列	4	-
過去購入額	数値	343	-
継続回数	数値	10	3%
メアド登録	文字列	2	-
1月以内ログイン	文字列	2	-
契約担当者名	数値		
継続・退会(ターゲット)	文字列	2	-

学習設定の情報を閲覧するには

学習設定情報をご覧ください。閲覧したいモデルをモデル一覧で選択すると選択したモデルに関する情報が表示されます。

学習時の項目に関する情報を閲覧するには

項目情報をご覧ください。閲覧したいモデルをモデル一覧で選択すると選択したモデルに関する情報が表示されます。入力項目として利用しなかった項目は薄いグレーの文字色になります。

予測画面

この画面はメイン画面で、予測タブをクリックすると表示されます。



新規に予測を行うには

モデル一覧で予測を行いたいモデルを選択した状態で、予測タブ内の新規予測ボタンをクリックします。

予測の履歴を閲覧するには

予測履歴ををご覧ください。過去に予測を実行した数だけ履歴が表示されます。

予測の履歴を削除するには

予測履歴のファイル名の右のチェックボックスをチェックし、削除ボタンをクリックします。

コマンドラインで予測を実行するには

コマンドではモデルが格納されているフォルダを指定します。フォルダ名は新規予測ボタンの下の文に記載があります (例: [ワークスペース]\20190524_224446.471\20190525_204851.975)。コマンドの他の部分については、本マニュアルの「コマンドライン機能」を参照してください。

予測設定画面

予測画面で新規予測ボタンをクリックすると、この画面に遷移します。



予測データを指定 (ファイルを指定) して、予測を実行するには

中央のウィンドウの領域にファイルをドラッグすることで指定することができます。別の方法として、ファイル指定ボタンをクリックし、ファイルブラウザから指定することもできます。ファイルフォーマットは CSV (カンマ区切り) と TSV (タブ区切り) に対応しています。

予測データの項目について、選択している予測モデルを作成した際の学習データと、予測ターゲット項目を除いて一致している必要があります。

ファイル指定後、決定ボタンをクリックすると予測を実行します。

予測結果のファイルに予測理由を追加するには

予測理由チェックにチェックをします。多値分類では予測結果に予測理由を付与することはできないため、チェックできません。

予測結果のファイルに予測データの内容を追加するには

予測データ追加チェックにチェックをします。チェックすると予測結果に対して予測データを追加します。

予測結果のファイルに予測データの特定の項目を追加するには

項目プルダウンから、追加したい項目を選択します。選択された項目が予測結果に追加されます。何も選択しない場合は行番号 (予測データファイルの何行目か) が表示されます。

予測プレビュー画面

予測設定画面でファイルを指定して決定ボタンを押すところの画面に遷移します。

予測結果プレビュー 予測結果を保存ボタン

完了

先頭10行のプレビュー

行番号	(a)退会	(b)継続	(a)退会の確率を上げる1	(a)退会の確率を上げる2	(a)退会の確率を上げる3	(b)継続の確率を上げる1	(b)継続の確率を上げる2
1	0.204763	0.795237	年齢:46:0.0235062	性別:未選択:0.0139136	入会時期:2011-7-22:0.0	顧客ランク:プラチナ:0.0	契約タイプ:契約タイプ:0.0
2	0.370848	0.629151	1月以内ログイン:なし:0	年齢:47:0.104679	過去購入額:7950円:0.0	メアド登録:あり:0.0313	継続回数:3.0:0.042369
3	0.605211	0.394789	顧客ランク:ブロンズ:0.0	1月以内ログイン:なし:0	過去購入額:20100円:0.0	メアド登録:あり:0.0384	性別:男性:0.00589906
4	0.235126	0.764874	顧客ランク:ゴールド:0.0	入会時期:2011-11-4:0.01	年齢:38:0.00360892	継続回数:3.0:0.042369	1月以内ログイン:あり:0
5	0.267365	0.732635	過去購入額:21540円:0.0	契約タイプ:契約タイプ:0.026	継続回数:10.0:0.016900	1月以内ログイン:あり:0	年齢:36:0.0313204
6	0.184019	0.815981	メアド登録:あり:0.0022	入会時期:2014-4-13:0.0	継続回数:6.0:0.001920	顧客ランク:プラチナ:0.0	年齢:36:0.0313204
7	0.186079	0.813921	1月以内ログイン:なし:0	契約タイプ:契約タイプ:0.000	継続回数:10.0:0.00534	顧客ランク:プラチナ:0.0	年齢:36:0.0313204
8	0.246309	0.753691	年齢:36:0.0313204	入会時期:2016-11-24:0.0	継続回数:1.0:0.013121	過去購入額:36000円:0.0	1月以内ログイン:あり:0
9	0.186806	0.813194	1月以内ログイン:なし:0	継続回数:10.0:0.006122	性別:女性:0.00589906	顧客ランク:プラチナ:0.0	入会時期:2016-2-5:0.006
10	0.533117	0.466883	顧客ランク:シルバー:0.0	1月以内ログイン:なし:0	入会時期:2016-2-5:0.006	メアド登録:あり:0.0491	性別:女性:0.00589906

予測結果を保存

予測結果のプレビューを閲覧するには

予測結果プレビューに、予測結果としてファイルに出力される内容の上位 10 行が表示されます。

予測結果を保存するには

予測結果を保存ボタンをクリックして、保存してください。

予測ファイルの説明

以下では、予測結果として出力されるファイルについて説明します。予測結果は CSV ファイルとして出力されます。

二値分類 (予測理由なし)

1 列目に予測ファイルと対応する行番号、2 列目・3 列目に予測ターゲット項目の 2 つの値に対する予測確率を表示します。

行番号	故障	正常
1	0.737798	0.262202
2	0.579275	0.420725
3	0.168248	0.831752

二値分類 (予測理由あり)

1～3 列目は予測理由なしの場合と同じです。4～6 列目は 1 つ目の値の予測確率増加に寄与した項目を 3 つまで、7～9 列は、もう一方の値の予測確率増加に寄与した項目を 3 つまで表示します。

予測に寄与した項目は、[項目名]:[項目の値]:[寄与スコア]のフォーマットで記述されます。寄与度スコアは大きいほど、予測確率増加への寄与が大きいことを示します。

行番号	故障	正常	故障の確率を上げる 1	・・・省略・・・	正常の確率を上げる 3
1	0.823877	0.176123	稼働月数:80:0.278897		
2	0.566978	0.433022	稼働月数:74:0.216476		圧力:100.01:0.002017
3	0.0339	0.9661	湿度:89.75:0.005551		圧力:101.42:0.000479

多値分類

1 列目に予測ファイルと対応する行番号、2 列目以降に予測ターゲット項目の各値に対する予測確率が出力されます。

数値予測 (予測理由なし)

1 列目に予測ファイルと対応する行番号、2 列目に予測値が出力されます。

行番号	予測値
1	99.00743
2	97.92587
3	94.79516

数値予測 (予測理由あり)

1 列目に予測ファイルと対応する行番号、2 列目に予測値が記述されます。3～5 列名は予測値増加に寄与した項目を 3 つまで、6～8 列は、予測値減少に寄与した項目を 3 つまで表示します。

予測に寄与した項目は、[項目名]:[項目の値]:[寄与スコア]のフォーマットで記述されます。寄与度スコアは大きいほど、予測確率増加への寄与が大きいことを示します。

行番号	予測値	予測値を上げる 1	・・・省略・・・	予測値を下げる 3
1	99.00743	日付:2018/2/1:755.9		
2	97.92587	曜日:月:1231.4		月末フラグ:1:24.5
3	94.79516	前年出荷数:18963:8273.2		曜日:日:311.1

予測データを追加した場合

予測結果ファイルの右側に予測ファイルの項目が連結されます。下記は二値分類の例です。4 列目以降に予測ファイルの項目が連結されます。

番号	故障	正常	顧客 ID	・・・省略・・・	契約担当者名 (仮)
1	0.823877	0.176123	ID02001		ノムラ マナブ
2	0.566978	0.433022	ID02002		コバヤシ トモミ
3	0.0339	0.9661	ID02003		サカモト トモヤ

項目を追加した場合

行番号の代わりに選択した項目が追加されます。下記は二値分類の例です。

顧客 ID	故障	正常
ID02001	0.823877	0.176123
ID02002	0.566978	0.433022
ID02003	0.0339	0.9661

コマンドライン機能

Prediction One は GUI ツールですが、一部の機能はコマンドライン機能を利用することで、GUI を使わずに実行することが可能です。

学習 (learn) コマンド仕様

学習データから予測モデルを生成するコマンドです。

PredictionOneCmd.exe **learn** [-md model_dir_path] [-od output_dir_path] [-lf learning_data_file_path] [-tc target_column_name] [-tt task_type] (-tv target_value) (-ef evaluation_data_file_path) (-ci column_info_file_path) (-cv cross_validation_k_fold)

オプション名	説明						
-md	モデルディレクトリのパス。実行するとモデルファイル群が生成される。						
-od	アウトプットディレクトリのパス。ディレクトリ内に以下のファイルが出力される。 <table><tr><th>ファイル名</th><th>説明</th></tr><tr><td><i>metrics.csv</i></td><td>評価データに対する評価結果。評価指標名と評価値。</td></tr><tr><td><i>analysis_summary.csv</i></td><td>評価データに対する項目の寄与度。</td></tr></table>	ファイル名	説明	<i>metrics.csv</i>	評価データに対する評価結果。評価指標名と評価値。	<i>analysis_summary.csv</i>	評価データに対する項目の寄与度。
ファイル名	説明						
<i>metrics.csv</i>	評価データに対する評価結果。評価指標名と評価値。						
<i>analysis_summary.csv</i>	評価データに対する項目の寄与度。						
-lf	学習データファイル(CSV/TSV)の入力パス。						
-tc	予測ターゲットとして使用する項目の名前。						
-tt	予測タイプ。二値分類なら <i>binary_classification</i> 、多値分類なら <i>multi_classification</i> 、数値予測なら <i>regression</i> を指定。						

-tv	予測値。予測タイプで二値分類を指定した場合は必須。それ以外は指定なし。
-ef	評価データファイル(CSV/TSV)の入力パス。省略した場合、学習データから自動抽出。
-ci	項目情報ファイル(CSV)の入出力パス。省略した場合、学習データの先頭 1000 行から自動で判定される。 項目情報ファイルのフォーマットは、項目情報生成コマンドの仕様 (-ci の説明部分) を参照。
-cv	k 分割交差検証の分割数。0 を指定すると学習データに基づき、交差検証を行うかどうかとその分割数を自動決定する。1 を指定すると交差検証を行わない。2 以上を指定すると、指定された数 k で k 分割交差検証を行う。省略した場合は 0 指定と同じ。

コマンドの例

PredictionOneCmd.exe learn -md model -od output -lf 二値分類_故障予測.csv -tc 状態(ターゲット) -tt binary_classification -tv 故障

予測 (predict) コマンド仕様

予測モデルを予測データに適用し、予測結果を生成するコマンドです。

PredictionOneCmd.exe **predict** [-md model_dir_path] [-od output_dir_path] [-pf prediction_data_file_path]

オプション名	説明	
-md	モデルディレクトリのパス。学習コマンドで生成されたモデルファイル群を含むディレクトリを指定する必要がある。	
-od	アウトプットディレクトリのパス。ディレクトリ内に以下のファイル名で予測結果が保存される。	
	ファイル名	説明

	<i>prediction_valid.csv</i>	正常に予測された行の行番号と予測結果。
	<i>prediction_analysis.csv</i>	正常に予測された行の行番号と予測結果と予測理由。(二値分類と数値予測の場合のみ対応)
	<i>prediction_skip.csv</i>	予測がスキップされた行の行番号。 (列数が異なるなど)
	<i>prediction_invalid.csv</i>	予測に失敗した行の行番号。(モデルとの不整合など)
-pf	予測データファイル(CSV/TSV)の出力パス。	

コマンドの例

PredictionOneCmd.exe predict -md model -od output -pf 二値分類_故障予測（予測用）.csv

項目情報生成(mkcolinfo)コマンド仕様

学習時に利用する項目情報ファイルを自動生成するコマンドです。自分で項目情報ファイル生成する際は実行する必要はありません。

PredictionOneCmd.exe **mkcolinfo** [-lf learning_data_file_path] [-ci column_info_output_file_path] (-tc target_column_name)

オプション名	説明
-lf	学習データファイル(CSV/TSV)の入力パス。
-ci	項目情報ファイル(CSV)の出力パス。 項目情報ファイルのフォーマットは以下の通り。

	項目名	説明
	<i>column_name</i>	項目名。学習データファイル中の項目名と完全一致している必要がある。
	<i>column_type</i>	項目の型。整数なら integer 、実数なら float 、文字列なら string 、日付なら date 、使用しない項目なら ignore 。
-tc	予測ターゲットとして使用する項目の名前。省略した場合、項目の型 (integer/float/string/date) 判定結果のみをそのまま出力する。指定した場合は、項目の型判定に加えて使用しない項目(ignore)の判定までを行った結果を出力する。	

コマンド実行例

PredictionOneCmd.exe mkcolinfo -lf 二値分類_故障予測.csv -ci column_info.csv -tc 状態(ターゲット)

FAQ

Q：データ作成の際に、取得できない値がある場合は (欠損がある場合は) ？

A：欠損値として空文字を書いてください。どのデータタイプでも同様です。

Q：数値を表す項目について、空文字とゼロは同じ扱いになるか？

A：いいえ、なりません。ゼロは値として扱われます。空文字は数値ではなく、数値を表さない特別な情報として扱われます。数値の項目に空文字以外を利用すると、項目が文字列タイプと判定されてしまうことがあります。

Q：年月日を表す項目のフォーマットは？

A：下記のどれかであれば、年月日として処理されます。

- "dd-MM-yyyy"
- "yyyyMM"
- "yyyyMMdd"
- "ddMMyyyy:HH:mm:ss"
- "yyyy-MM-dd"
- "yyyy/MM/dd"
- "ddMMMyyyy:HH:mm:ss"

d は日を、M は月を、y は西暦を、H は時間を、m は分を、s は秒を表します。

Q：データファイルの文字コードは？

A：UTF-8 (BOM あり) と Shift_JIS に対応しています。

Q：データファイルの文字コードは Shift_JIS なのに文字化けが発生するのはなぜ？

A：Prediction One ではファイルの先頭 1000 バイトを読み込んで文字コードを判定しています。そのため、先頭 1000 バイトに文字コードを判定できる情報がないと誤判定が起きることがあります。その場合、データファイルを UTF-8 (BOM あり) に変換すると、文字化けがほぼ起きなくなります。

Q：ツールでファイル読み込み後に、ファイルを編集したらどうなる？

A：予測不能なエラーが起きる場合が多いので、ファイルを編集したあとは読み込み直してください。

Q：学習を開始して「データ読み込み」の状態で数分経っても進まない。

A：学習データのサイズが大きいと、学習中の各処理にかかる時間は増加します。学習データのサイズが大きい場合、そのままお待ちください。

Q：学習データとして使用可能なデータの規模(行数・列数)は？

A：100 行～100 万行、2 列～200 列のデータを用意してください。行数や列数が増えるほど、学習にかかる時間・メモリ使用量は増加します。メモリ使用量がお使いの PC の容量を超えるとソフトウェアが終了する場合があります。

Q：二値分類と多値分類で、予測ターゲット項目の出現頻度の偏りが大きいデータセットで学習すると「学習が失敗しました」というエラーが表示されるのはなぜか？

A：学習データのみを指定し、評価データを指定しない場合、Prediction One は内部を学習データと評価データに分割します。分類では、学習データや評価データの予測ターゲット項目に 1 種類の値しか出現なくなると、学習ができなくなります。予測ターゲット項目に複数の値を含めるように評価データを作成し、評価データとして指定すると、この問題を回避できます。

Q：インストール中に「使用中のファイル：更新する必要があるファイルが使用中です」と表示され「自動的に閉じて、アプリケーションを再起動する。」を選択し「OK」をクリックすると、Windows のエクスプローラーが再起動せずタスクバーやスタートメニューが表示されなくなったのはなぜか？

A：お使いの PC の環境やインストール済みのソフトウェアによって、アプリケーションが正しく再起動されないことがあります。エクスプローラーを再起動するには、Ctrl+Alt+Del キーを同時に押してタスクマネージャーを起動し、「ファイル」→「新しいタスクの実行」に explorer.exe と入力して「OK」をクリックしてください。

Q：サンプルデータで「予測用」のファイルの評価データに指定すると評価結果が異常となるのはなぜか？

A：予測用のサンプルデータは予測ターゲット項目が全て「?」となっているため、評価データとしては使用できません。評価データとして使用するには、予測ターゲット項目に実際の値が正しく入力されているファイルをご用意ください。

Appendix: 評価値に関する参考資料

分類の代表的な評価値： accuracy, precision, recall, F値

予測結果とクラス

予測確率	予測	実際の結果
10%	正常	正常
88%	故障	故障
21%	正常	故障
⋮		⋮
98%	故障	正常

例えば、予測確率>50%なら故障と予測

4つのパターンの数をカウント (混同行列)

	実際に故障	実際に正常
故障と予測	200	50
正常と予測	100	1000

$$accuracy = \frac{1000 + 200}{1000 + 100 + 50 + 200} = 88.8\%$$

$$precision = \frac{200}{200 + 50} = 80.0\% \quad \text{検出時の正確度}$$

$$recall = \frac{200}{200 + 100} = 66.6\% \quad \text{検出できた割合}$$

$$f \text{ value} = \frac{2 \cdot recall \cdot precision}{recall + precision} \quad \text{precisionとrecallの総合力}$$

数値予測の代表的な評価値

予測価格 x	実際の価格 y	絶対誤差 x - y	2乗誤差 (x - y) ²	絶対値誤差率 x - y /y
800万	1200万	400万	400万 ²	33.3%
4600万	5000万	400万	400万 ²	8.0%
3000万	6000万	3000万	3000万 ²	50.0%
⋮				⋮
3000万	3200万	200万	200万 ²	6.25%

- 絶対誤差よりも2乗誤差の方が外れ値 (通常より大きな値) に敏感
- 値の誤差よりも誤差率の方がふさわしい場合もある
- 外れ値が多い場合、平均値よりも中央値の方が適切な場合もある

平均
絶対誤差

$$\frac{1}{N} \sum_{i=1}^N |x_i - y_i|$$

RMSE
Root Mean
Squared Error

$$\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2}$$

誤差率
中央値

$$\frac{1}{N} \sum_{i=1}^N \frac{|x_i - y_i|}{y_i}$$

(データインデックスを i、データ数を N とする)

$$\text{決定係数: } 1 - \frac{\sum_{i=1}^N (x_i - y_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$$