rapidminer
KSK ANALYTICS JAPAN PARTNER

Auto Model のご紹介

RapidMinerを使ってデータ分析を始めるのか、その他の古い手法で行うかどうかに関わらず、オートモデル（Auto Model）は普段のモデル作成業務をより簡単に、そして効率化することができます。オートモデル（Auto Model）は RapidMiner Studio の有償の拡張機能であり、「プロセスの構築」や「モデルの評価・検証」の構築を加速させます。最も優れていることの一つとして、自動で作成したモデルを自身の手で「修正」することも可能であり、作成したモデルに手を加えることができます。作成モデルがブラックボックス化することを回避します。

オートモデルは分析課題に対して、3 つのアプローチを用意しています。

・Prediction(予測)

・Clustering(クラスタリング)

・Outliers(外れ値)

予測のカテゴリーでは、分類問題と回帰問題の両方を解決することができます。とりわけ、教師あり学習によるモデル作成を行う場合に使用します。オートモデルはデータを評価することを助け、問題解決のための関連モデルを提供し、モデルの結果を比較・検討することの手助けを行います。

オートモデルは結果を得る事を助けるだけでなく、ディープラーニングのようなおそらく理解することが困難である内部のロジックの結果を理解することも助けます。具体的には重要な説明変数の可視化や作成されたモデルに基づく、予測値のシミュレーションを行うことも可能です。

クラスタリングのカテゴリーでは、主に教師なし学習によるモデル作成を自動化します。正常な振動データしかない場合、同じ正常でもいくつかのパターンが存在することを確かめたい場合などに用います。クラスタリングによりグループ化された各グループの特徴を可視化する機能も備えているので、グループの特徴を考察するのに役立ちます。

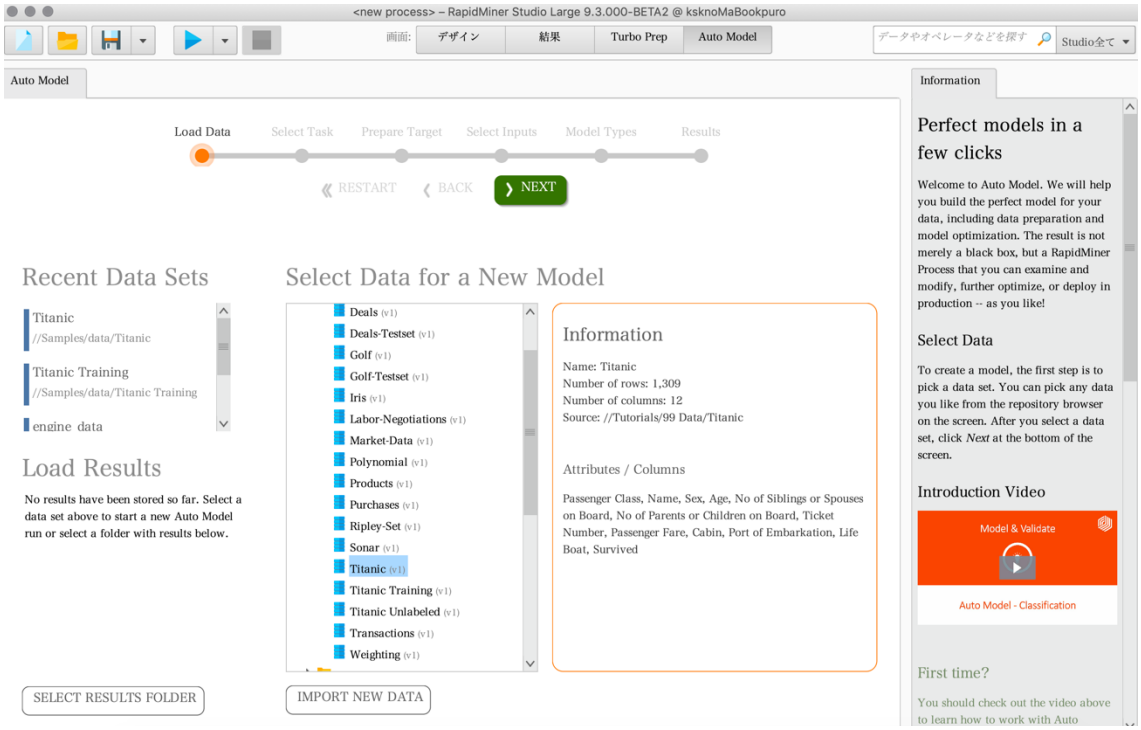
外れ値のカテゴリーでは、外れ値の抽出（Outlier Detection）を行います。グループ内の特定の個体に他とは異なる挙動がみられた際に外れ値データ(Outlier)として検出することができます。例えば、Web サーバー群の特定の 1 サーバーが異常なリクエスト数を処理しているような場合を検出し、これをリプレースすべきか判断することができます。

例題：タイタニック号での生存予測モデル

オートモデルをどのように使うのか具体的なイメージをお伝えするために、以下では、タイタニック号での生存者のデータセットを使って、自動モデル作成を行います。開始するために、RapidMiner Studio の一番上にあるボタンを押してオートモデルビューを選択します。

データ選択

オートモデルへの最初のステップはリポジトリにあるデータセットのうちの一つを選択することです。もし、リポジトリの中に使用したいデータがなければスクリーン一番下の“import new data” と書いているリンクをクリックします。今回の例では、タイタニック号のデータセットは sample→data の順に開けた中に見つけることができます。このデータセットを選択し、そしてスクリーン下部にある Next ボタンをクリックします。

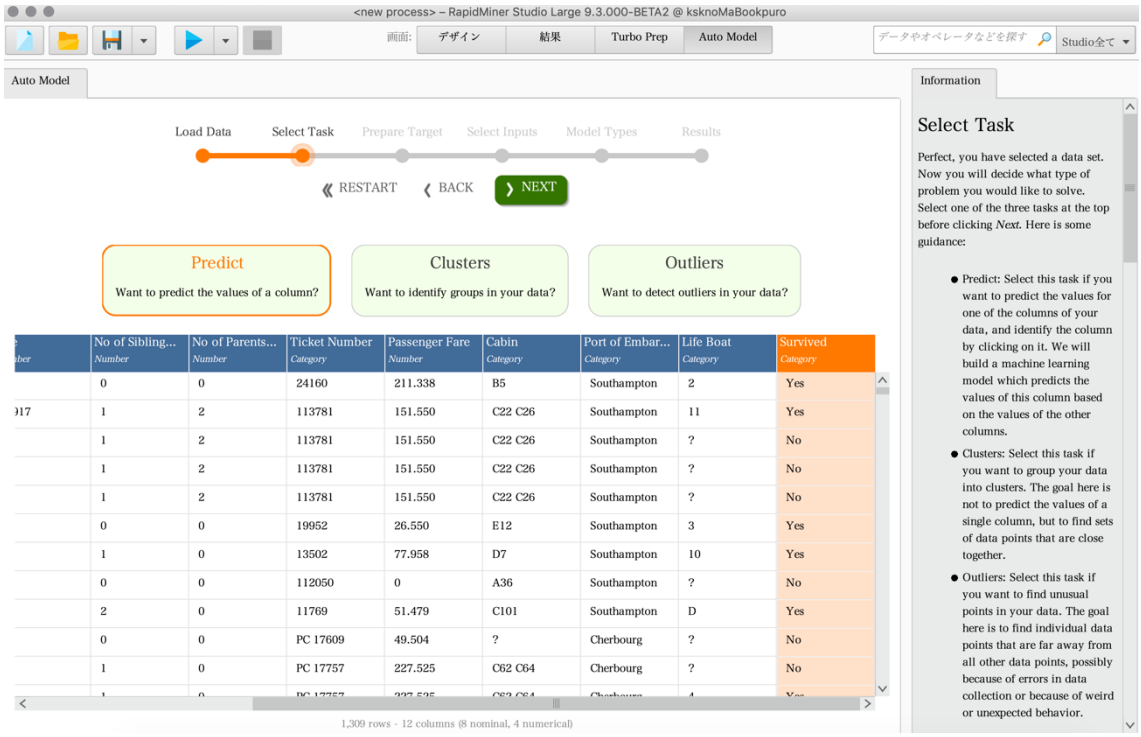


タスクを選択

データセットを選択し終えたら、どんな種類の問題を解決したいのかを決める必要があります。オートモデルでは 3 つの異なったタスクの中から目的のタスクを選択します。

・予測する(列の値を予測したい?)
・クラスタリング(データの中からグループを確認したい?)
・外れ値(データの中から外れ値を検出したい?)

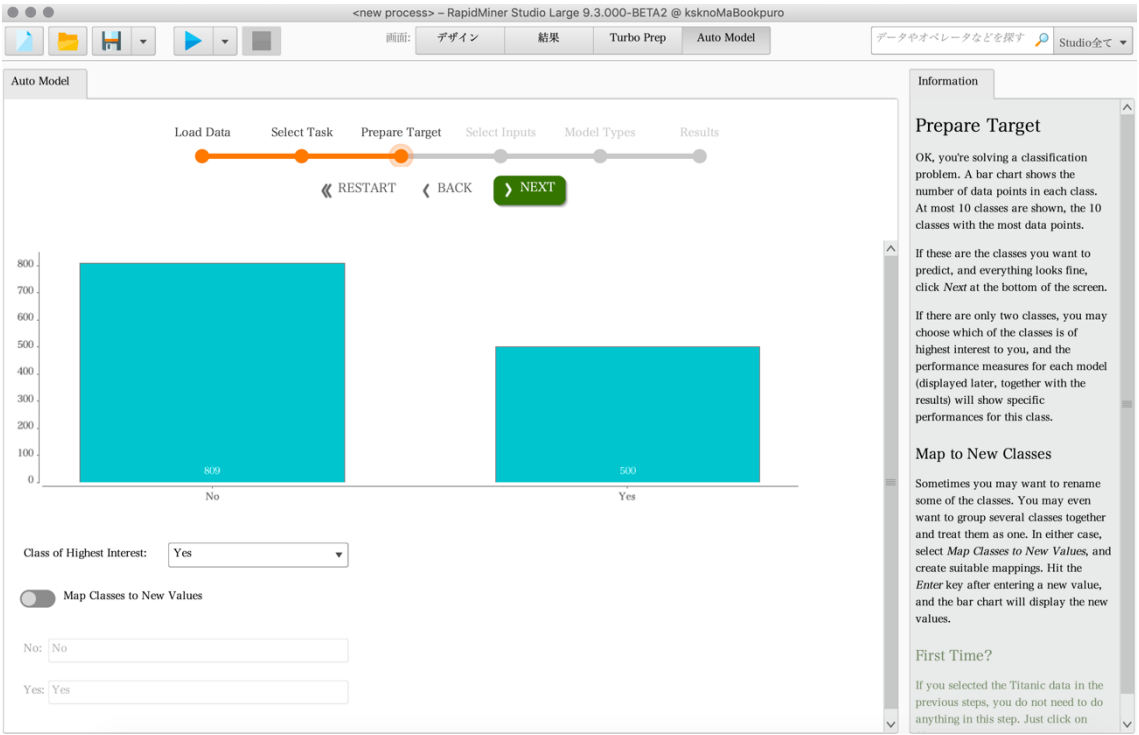
タイタニック号での生存者を予測したいので、今回のケースでは Predict を選択するべきです。そして Next をクリックする前に"Survived"列をクリックします。予測対象の列がオレンジ色になっていることを確認し、Next ボタンをクリックします。



ターゲットの準備

“Survived”は、”Yes”または ”No” の二つの値しか持っていないので、今回の問題は分類問題となります。たいていの場合は、分類問題のため、各クラスにデータ要素数を示す棒グラフを表示します。10 クラス以上の時は、最もデータ要素を持つ 10 クラスだけが表示されます。

ここで重要なことは、不均衡データでないかどうかを確認することです。オートモデルにより異常検知を行う場合に、正常データが全体の内 99%、異常データが全体の内 1%しかない場合があったとします。この場合、明らかな不均衡データと判断できます。不均衡データの処理を行った上で、モデル作成を行う必要があるため、ターゲットの準備段階で、予測対象列の不均衡具合が許容できるかどうかこの画面で検討を加えます。



・Class of Highest Interest

最も重要なクラスは後の結果を出力する時に重要になります。なぜなら、”Precision(精度)” と ”Recall(再現率)”のような性能値はどちらのクラスが“Positive(有用)”な結果かということに依存しているからです。今回のタイタニック号の例の中では、最も重要なクラスは”Yes”です。

・Map Classes to New Values

このステップは”Yes”と”No”からいくつか他の値に対して目標数値を命名することのオプションです。

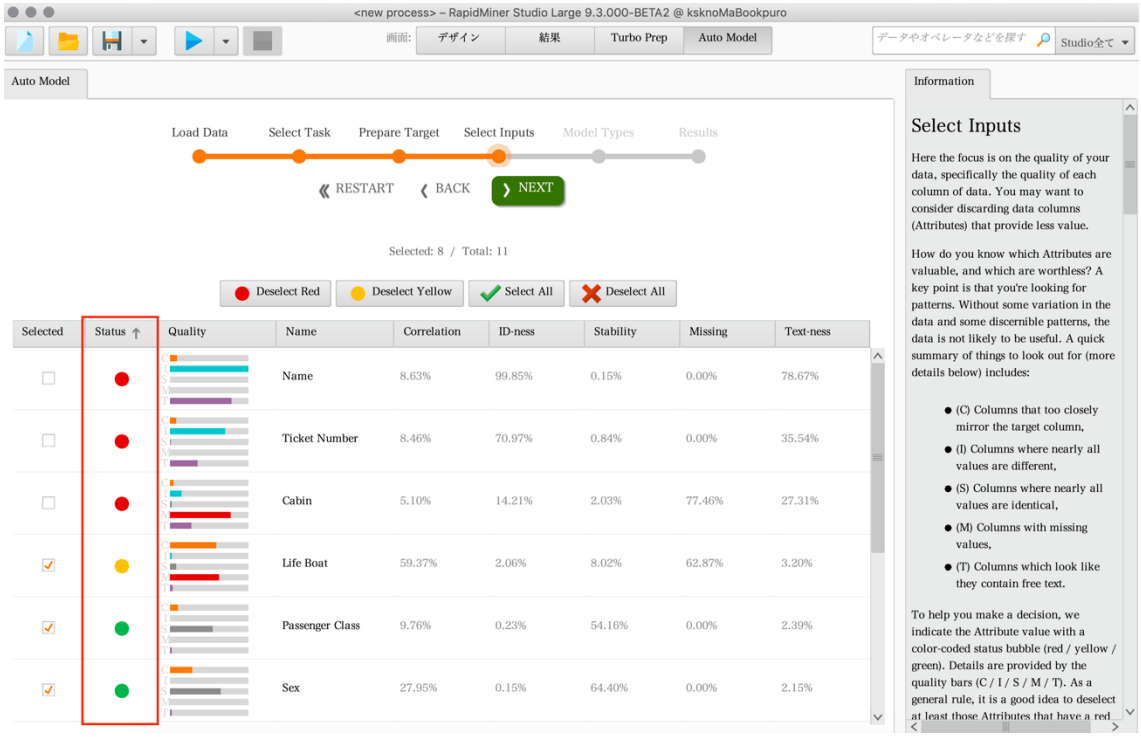
変数の選択

今回インプットされたデータ列の全てが予測を行うことを助けてくれるわけではありません。いくつかの列を取り除くことによってモデル作成のスピードを向上させ、パフォーマンスの改善に役立てることができます。しかし、どのようにしてその判断を下せばいいでしょうか？キーポイントは「パターン」を見つけ出すことにあります。対象データにおける変化といくつかの確認できるパターンを除いて、予測に役立ちそうにないデータを取り除きます。

取り除く必要のある列を見つけ出すために以下の 4 つの判断の根拠として、「赤」、「黄」、「緑」のステータスバブルを表示します。

・予測対象の列ととてもよく似ている列（予測対象列とかなり強い相関がある）
・ほとんど全ての値が異なる(例、ID など)
・ほとんど全ての値が同じ(一定性)
・欠損値を含んでいる(欠損値)

オートモデルでは以下の画面のように状況を色分けされたステータスバブルで要約します。一般的なルールとしては、少なくとも赤色のバブルの列を外すのは良い考えですが、黄色については自身で少し考えてみる必要があります。



【赤いステータスバブル】

タイタニック号の場合、“Name(氏名)”と“Ticket Number(チケット番号)”が ID に相当します。“Cabin (客室)”の値は最も多くの乗客で欠損しています。そのため、これら 3 つの行が赤いステータスバブルとなっており、モデル構築する際に除外する必要があります。パターンを発見することにおいて、赤いステータスバブル以外のものは役に立ちます。

【黄色のステータスバブル】

“Life Boat(救命艇)”は黄色のステータスバブルとなっていますが、それはこの列が今回の予測対象データである”Survived(生存者)”と強く関連しているからです。“Life Boat”と”Survived”は事実上同義語ですので、“Life Boat”行からデータを取り除き、そして生存者の基本的な情報からモデルを作成してみましょう。

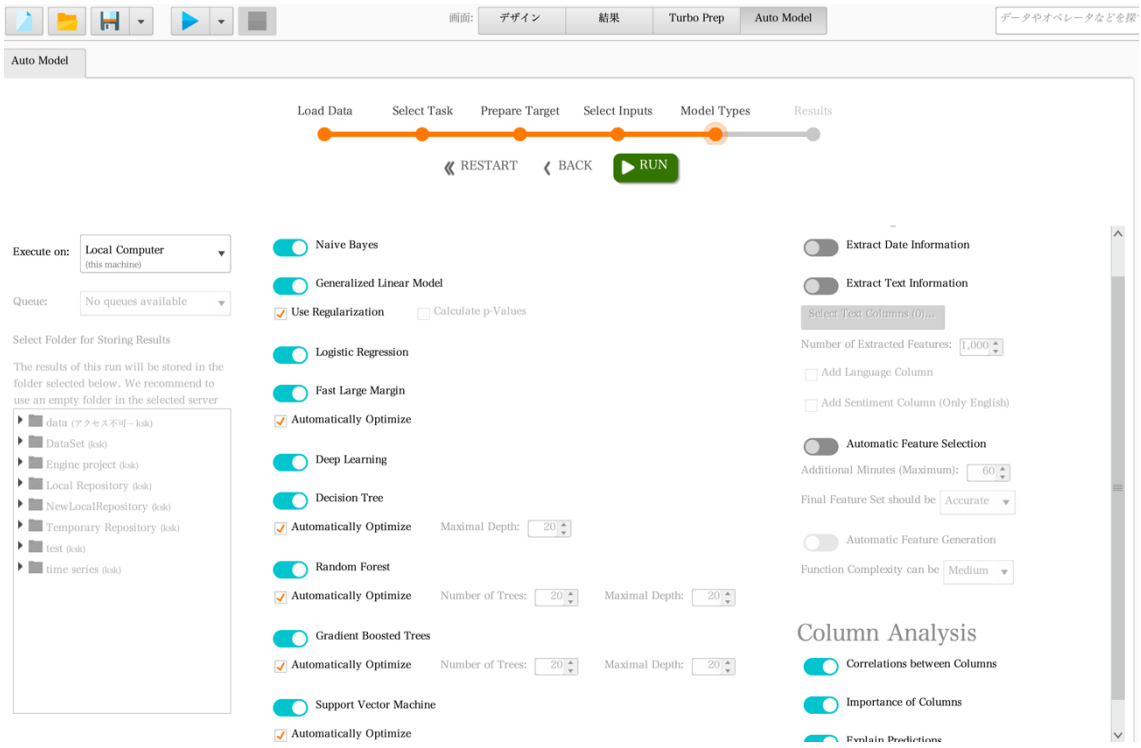
少し言い換えると、モデルが計画を立てるのに役立つかどうかの視点で作成を進めていくことです。乗客は先に救命艇に乗るつもりかどうかを知ることはできないので、計画時には役立ちません。しかし、チケットにどれくらいの料金を支払うかということや家族を同伴するかどうかということは、計画段階で知ることができます。

この例題では、あなたは黄色のステータスバブルを伴っている“Life Boat”を外すべきで、そして Next を押します。

モデルの選択

オートモデルは関連のあるモデル選択を提供してくれます。時間制約がある中でも、対象のデータセットに対して 9 つのアルゴリズムを使ってモデル作成、最適化を行います。モデル構築が終わるとパフォーマンスを比較し、ベストオプションを提示します。完成したモデルの正確性（Accuracy）または構築するのにかかる時間（Run Time）のどちらを優先するかは、結果をもとに自分自身で決めることになります。オートモデルでは合理的な妥協案にいち早く辿り着くことを助けます。

オートモデルは以下の 9 つのアルゴリズムを提供します。 データセットとアルゴリズムの相性を確認するために、線形、非線形、階層構造、木（ルール）のような様々なタイプのアルゴリズムが用意されています。



【用意されているアルゴリズム】
・ナイーブベイズ
・一般化線型モデル
・ロジスティック回帰
・ファストラージマージン
・ディープラーニング
・決定木
・ランダムフォレスト
・勾配ブースティング決定木(XGBoost)
・サポートベクタマシン

また、9.0 以降新たな実行機能も追加されるようになりました。例えば、上記の画面の右側にある「Automatic Feature Selection」、「Automatic Feature Generation」は、自動で変数選択（生成）を行う機能です。実行に時間を要するため、時間制限を指定するといいかもしれません。その下にある Column Analysis では、有効にすることで、変数間の相関、重要な変数、予測モデルに対する各変数の説明力をアウトプットしてくれます。

そのほかの新しい機能としては、実行速度を速めるために実行環境を選択することができます。ここでは、Local Computer を選択していますが、Server と接続していれば、Server 側で実行することも可能です。計算時間を短縮する場合は、Server 側で計算することになるでしょう。

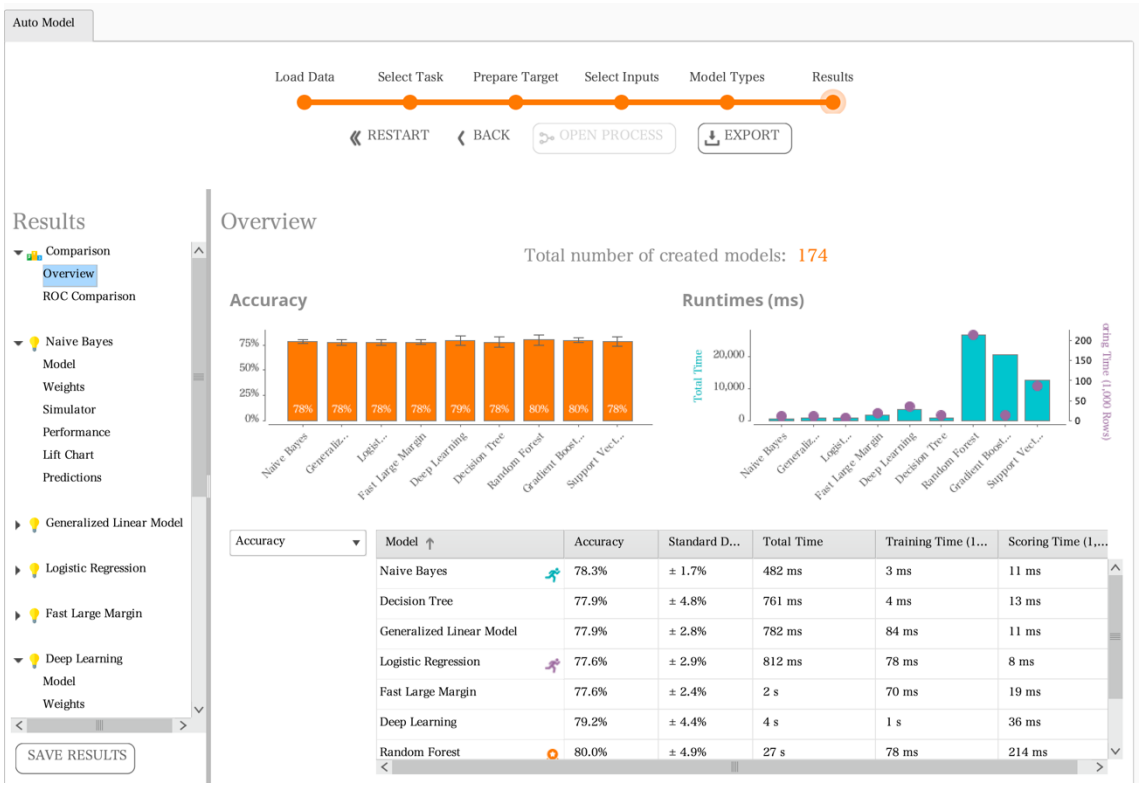
Run を押しモデルを構築し、結果を生成します。

モデルの結果

実行結果を得るまでしばしば時間を要します。最上部の Accuracy は各アルゴリズムの正答率を示しており、右側は各アルゴリズムの正答率を算出するまでに要した時間を意味しています。経過バーは進行中の計算のステータスを追跡しており、Stop ボタン押すことでモデル作成を止めることができます。

今回の結果を見ると、174 のモデルが作成され、上記 9 つのアルゴリズムの中で最も正答率が高かったもののみその結果が残されています。もちろん交差検証（Cross Validation）は実施済です。

今回のケースでは、ランダムフォレストの正答率が 80.0%と最も高く、計算時間という観点も加えるとナイーブベイズの計算時間が短い割に 78.3%と高い正答率が得られていることが見て取れます。一方で、決定木は木の深さ（Maximum depth）4 の場合が、一番正答率が高い（77.9%）という結果が得られましたが、決定木は他のアルゴリズムと比べると上手く予測できていないと言えます。



以下では、上記の画面の右側（Result）に表示される Weight、Performance について解説を行います。

【Weight】

Random Forest の Weights を見ると、重要な変数から順番に並んでいます。今回のデータセットでは、性別が重要な変数になっていることが確認できます。

Random Forest - Weights

Attribute	Weight
Sex	0.688
Passenger Class	0.121
Age	0.097
Passenger Fare	0.045
No of Siblings or Spouses on Board	0.040
Port of Embarkation	0.033
No of Parents or Children on Board	0.028

【Performance】

プログラムだと作成に手間がかかるクラス分類の結果を示す confusion matrix も自動で作成してくれます。左側の Criterion を選択することで precision、recall、F 値を得ることも可能です。

Random Forest - Performance

Criterion
- accuracy
- classification error
- AUC
- precision
- recall
- f measure
- sensitivity
- specificity

Table View / Plot View

accuracy: 80.00% +/- 4.90% (micro average: 80.00%)

	true No	true Yes	class precision
pred. No	193	42	82.13%
pred. Yes	33	107	76.43%
class recall	85.40%	71.81%	

その他、Lift Chart や Optimal Parameters などで作成した複数モデルの比較や結果の解釈を推進することができます。

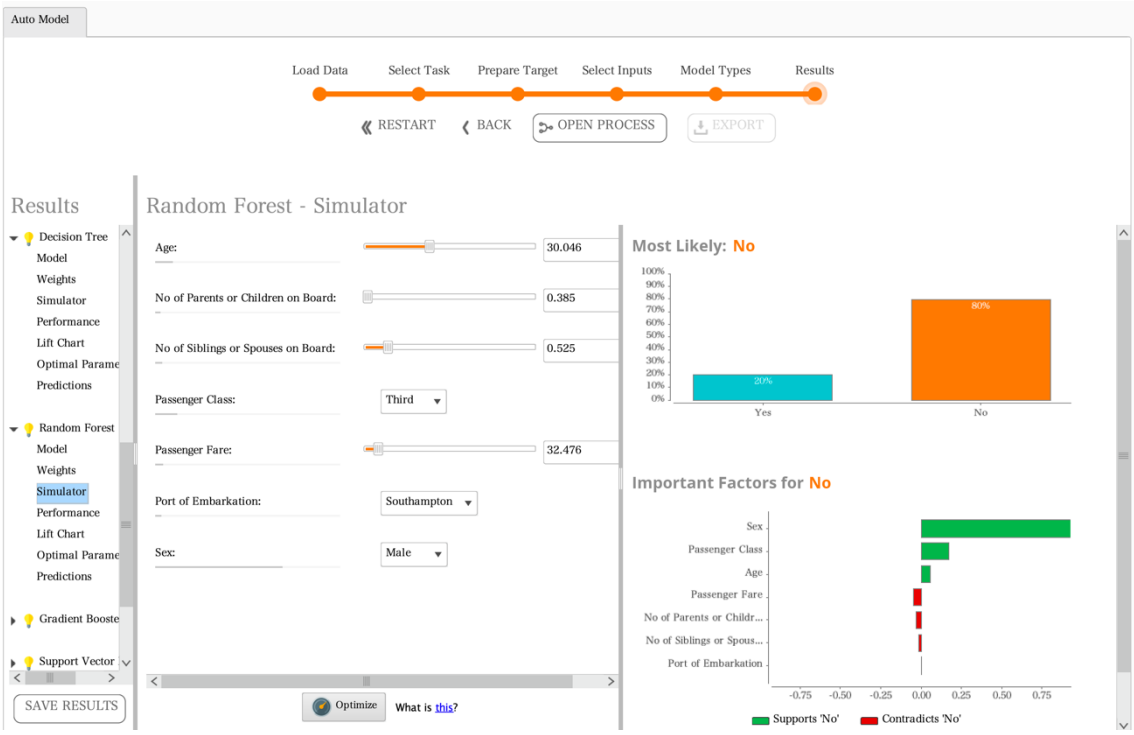
モデルシュミレーターとその他の有用なオペレーター

オートモデルは結果を得るための助けになるだけでなく、その結果を理解することも助けてくれます。ディープラーニングは計算式に当てはめ実行することは容易ですが、そのモデルの解釈は容易ではありません。下の画面において、オートモデルによって提供されるいくつかの有用なユーザーインターフェイスを使うことで、ディープラーニングによって算出される結果について探っていくことができます。

モデルシュミレーター

新たな知見を得るために、Deep Learning を開くと上から二番目に配置している Simulator を選択します。開くとスライダーとドロップダウンリストが左にあり、そして棒グラフが右にあるユーザーインターフェイスが見えます。最初の状態ではデータの平均の値が選択されています。タイタニック号のデータセットの場合、この平均はサードクラス、年齢は約 30 歳、男性で、相対的に少ない親戚が乗船している人々に相当します。

このシナリオで起こりそうなことは、右側にある上部の棒グラフによって、この乗客が生存しない確率が高いことが分かります。彼が生存する確率は 20%です。下部の棒グラフは何が彼に逆らっているか、とりわけそれは緑色のバーで示されている性別と乗客クラスです。この状況では性別と乗客クラスが生存者予想と一致していることを示唆しています。乗車運賃と船上の親戚を表している赤色のバーは予想と一致しないことを暗に示しています。



モデルシミュレーターの優れた点は双方向で、全ての値を変化させられることと、すぐさま予想の効果を見ることできることです。性別を男性から女性に変えてみると、およそ 50%生存確率が上昇します。それから、乗客クラスをファーストかセカンドに変えてみると生存確率が 90%以上となります。

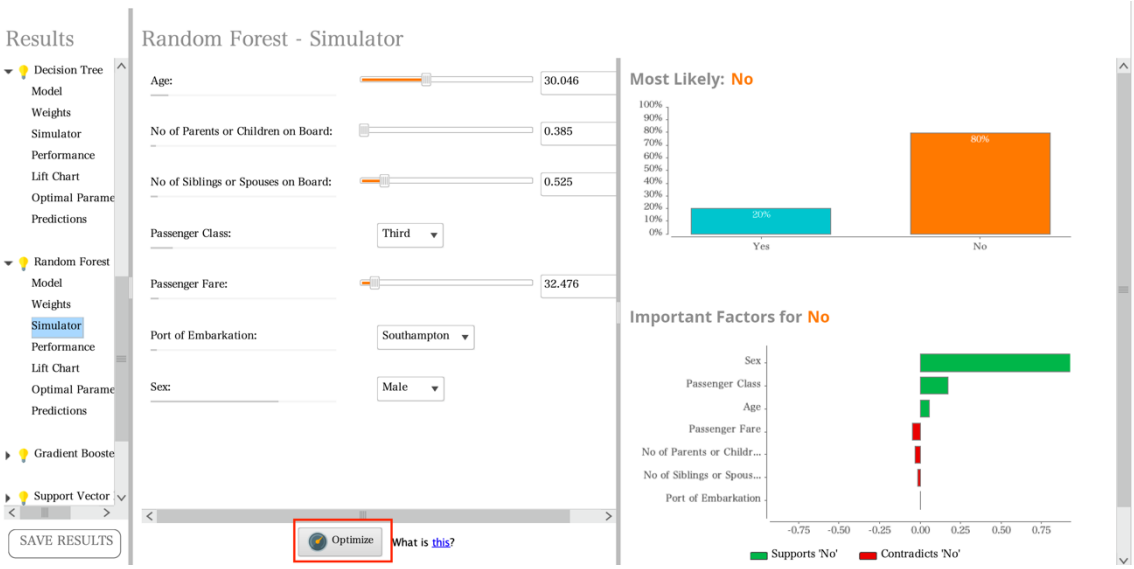
全てのスライダーとドロップダウンリストを操作することで、ディープラーニングでモデル構築されたにも関わらず、素早くモデルのための直感を構築することができます。モデルシミュレーターは単一データ要素(局地的な相関)の近くでのモデルの行動を解析することによって予想ができるような仕組みになっています。全体的な重要な意味を持つデータ欄(全体的な相関)を見るために属性名の下に示されているグレーのスケールバーに目が留まります。このうち、最も長い棒は乗客クラスと乗客運賃に続く性別の下に見えます。

最適化

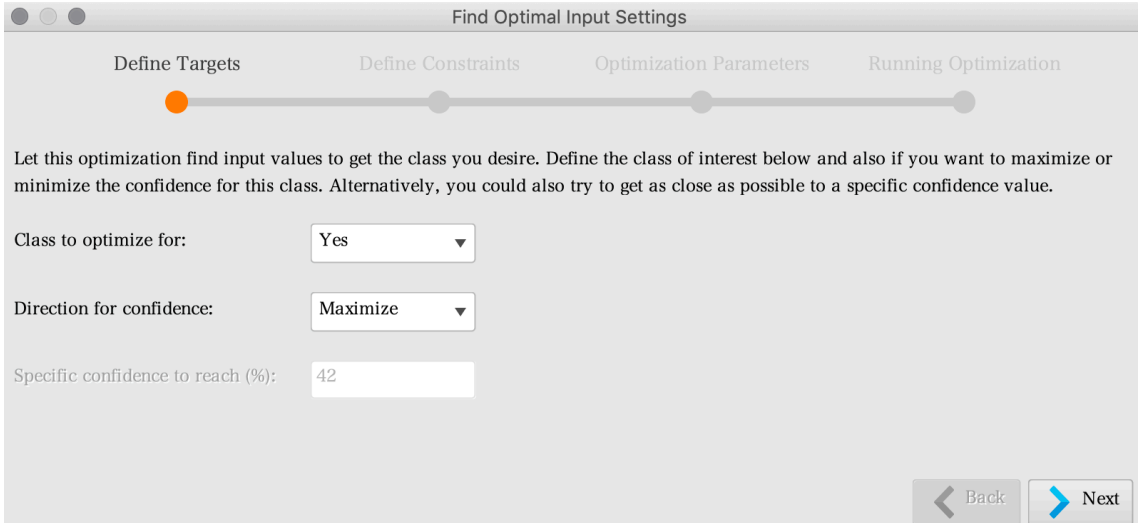
次の質問は、どのようにして乗客はタイタニック号での生存可能性を最適化することができますか?ということです。ここでも、オートモデルは答えを持っています。シミュレーターの左下部に、ラベルが付いた Optimize(最適化)ボタンがあります。

このボタンを押し、ダイアログの固定をするレシピ構築をサポートしてくれます。男性はタイタニック号上で女性よりも危険性があるので、男性のために生存する戦略を考えてみたいと思います。

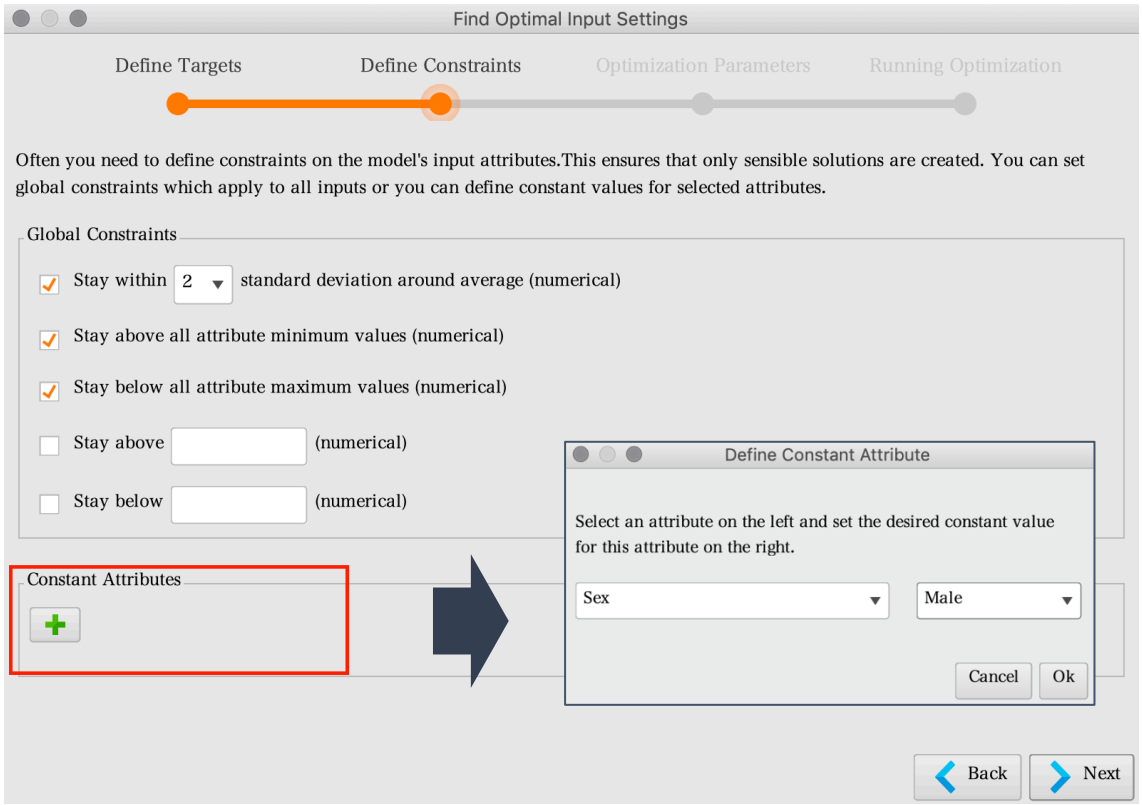
Optimize を押し、後に続くステップを試してみてください。



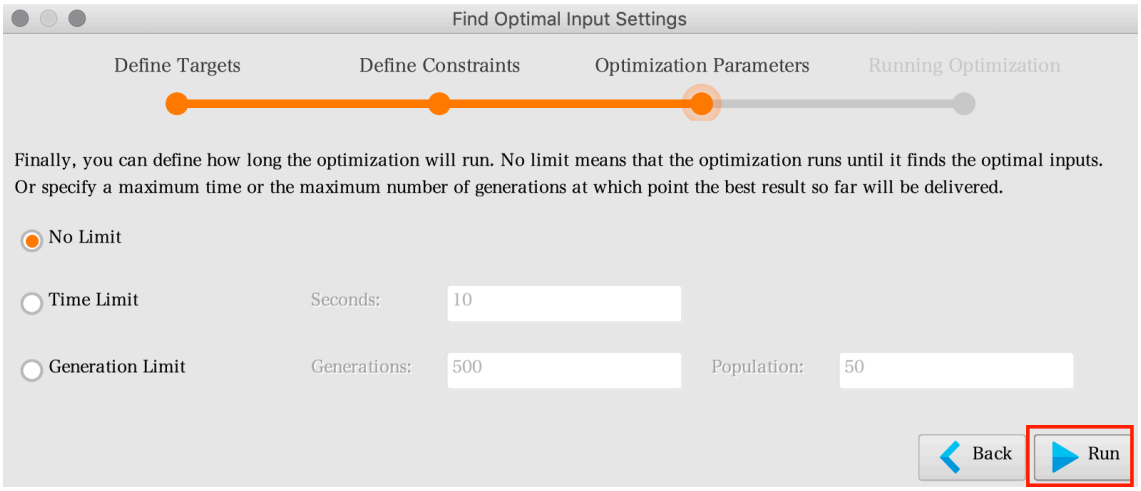
1 .Define targets の Class to optimize for において”Yes”を選択し、Next を押してください。



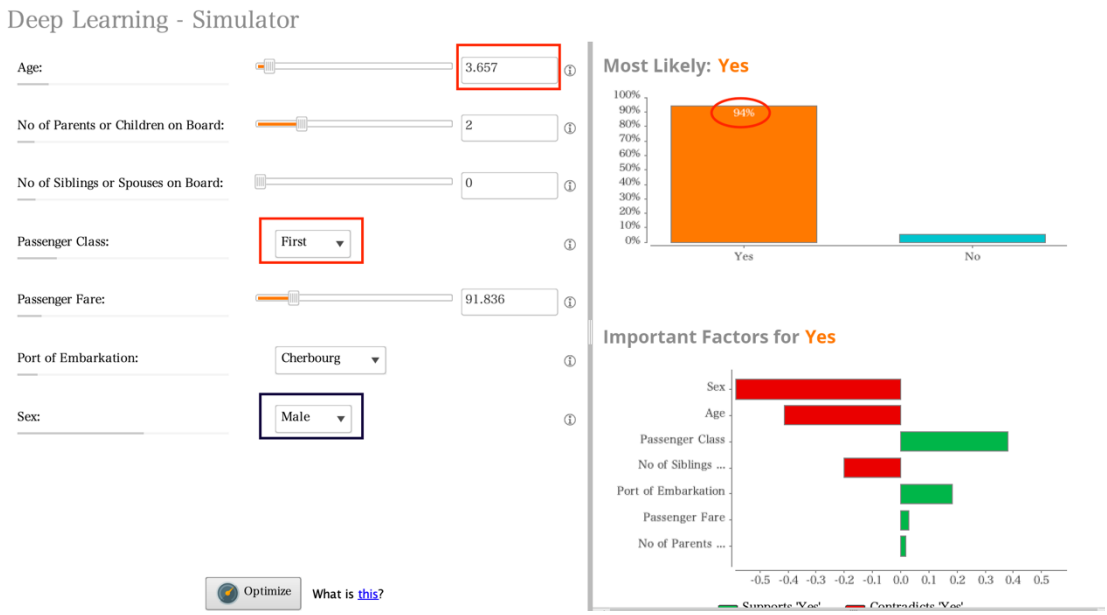
2 .左のドロップダウンリストの中から Sex(性別)を選び、右側のドロップダウンリストの中から Male (男性)を選択し、OK ボタンを押し、Constant Attributes に Sex = Male とあるのを確認し、Next ボタンを押してください。



3．Optimization Parameters において、Run を押してください。計算が完了すれば、Finish を押してください。



結果はシミュレーターの中にすぐに表示され、その結果は印象的です。男性の乗客で最も生存の望みがあるのは４ 歳で少数の親類とセカンドクラスで旅行をしている子どもです。彼の生存の可能性は 94%です。



タイタニック号に乗っている間、クラスは重要な変数となっていますが、ドロップダウンリストから乗客のクラスを変更すると分かる通り、サードクラスで旅行している 4 歳の男の子でさえも 70%を超える生存確率を有しています。

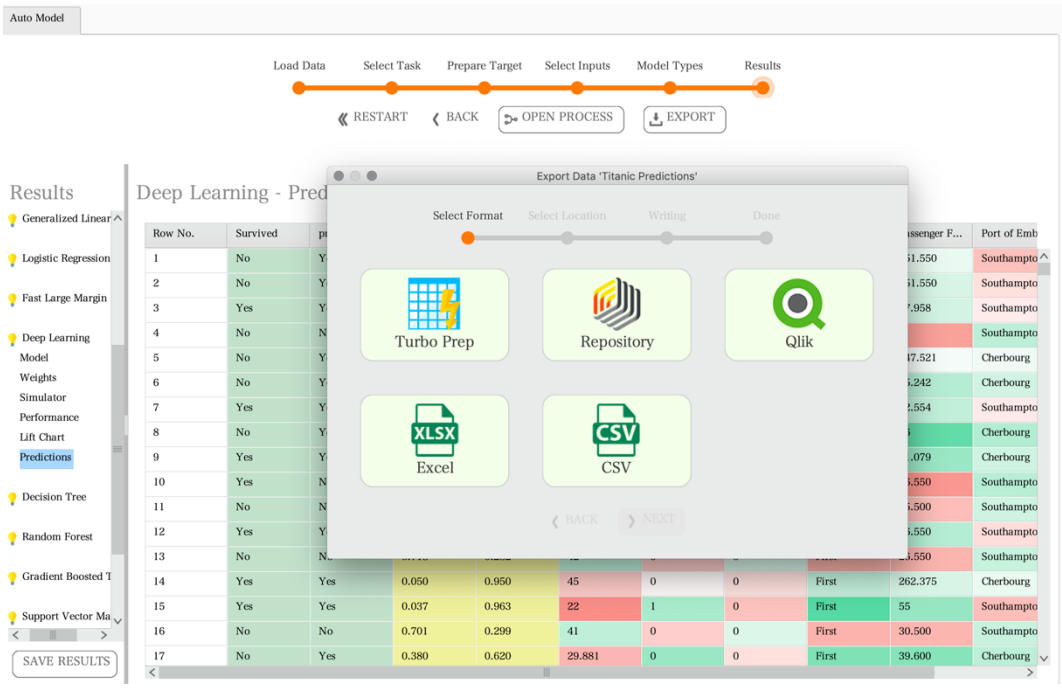
モデルシミュレーターはタイタニック号上で乗客がライフボートにおいて女性と子どもを優先の原理を思ったよりも厳格に固守したものと考えられます。Age(年齢)のスライダーを動かすことで、年齢を上げるにつれて生存確率が連続して減少することも確認できます。男性の乗客の生存確率が 50%を下回る年齢は、クラスに関連して変化します。

・サードクラス → 20 歳
・セカンドクラス → 29 歳
・ファーストクラス → 38 歳

より厳密に言えば、男性の乗客がどのようにしたら生存確率を高くなるかという質問には 答えることができません。年齢は与えられたものであり、より高い値段のチケットを買うことは金銭的な事情により叶わない場合もあるでしょう。しかし、最適化という方法とモデルシミュレーターはタイタニック号のデータから新たな知見を得ることを手助けしてくれます。

結果の出力

Predictions を選択することによりモデルに基づいた予測結果を出力することも可能です。予測結果を出力（格納）し、BI ツールなどで可視化することもできます。

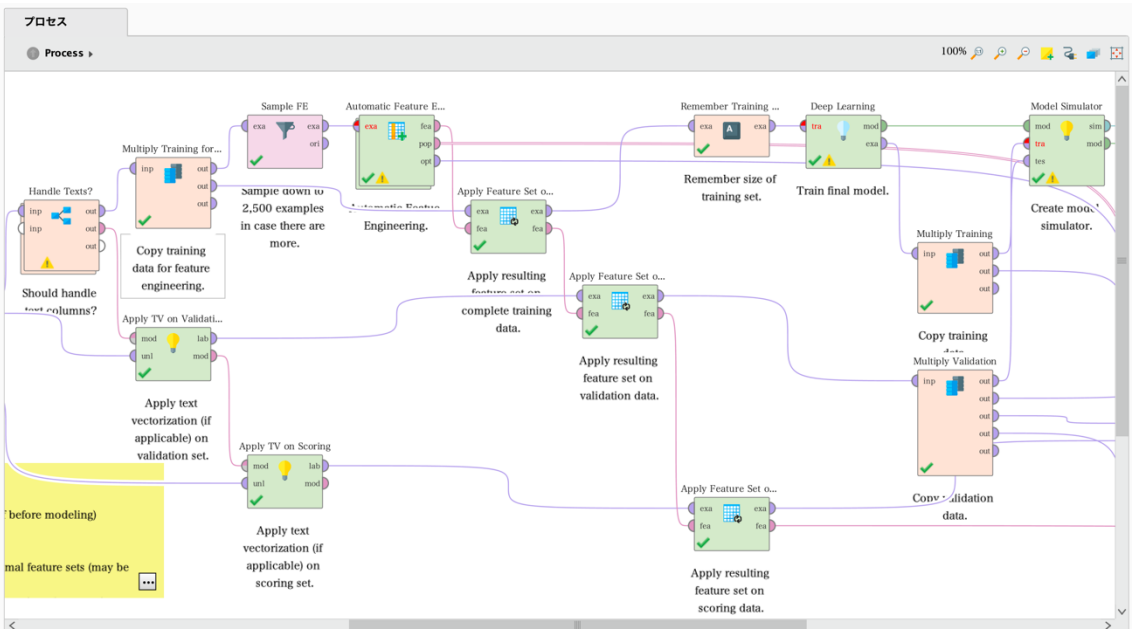


No Black Box !

オートモデルは有益なモデルを提供しますが、その分析過程を自らの手で確認し、修正したいと考えることがあるはずです。モデルシミュレータの下部にある Open Process を押すと、構築したモデルが RapidMiner のデザインビューに表示されます。このプロセスを実行することもでき、修正することもできます。オートモデルは問題解決するためのプロセスをブラックボックス化させません。

この点を強調するのには少なくとも 3 つの理由があります。

1. 中身の理解なくして、自信を持って製品の中にモデルを組み入れることはできないでしょう。どのようにモデルが機能し、全てが正しいと証明するのか確認したいはずです。
2. 新しいデータサイエンティストのプロセスを調べることで最も良い方法を学ぶことができます。
3. 巧みなデータサイエンティストは自身が作成するモデルのスタート地点として、オートモデルを使う事でその生産性を高めています。



近年は、MI（Material Informatics）の領域をはじめとして、研究所や技術開発室において、RapidMiner のオートモデルを使用し、実験データを活用する取り組みが活発化しています。もちろん、異常検知や予知保全、需要予測の領域においていくつもの活用例が報告されています。作成したモデルをベースに会議を進めれば、より有意義な考察が可能になるでしょう。紙幅の制約もあり、オートモデルのクラスタリングや外れ値の解説はできませんでしたが、別の機会に説明させていただければと思います。