＜　自習用コンテンツ紹介　＞

【Part1： RapidMiner ダウンロード】
　RapidMiner無償版のダウンロード方法

【Part2： データ読み込み】
　データを取り込む方法

【Part3： データ可視化】
　より良い知見を得るためのデータの可視化方法

【Part4： 予測モデル作成】
　タイタニック号事件の生存者を予測するモデルを作成する方法

【Part5： 予測モデルの評価】
　良い結果を保証するモデルの精度を評価する方法

※ 参考資料：
　KSK アナリティクス(株)　「RapidMiner7 Learn Tutorials」
　KSK アナリティクス(株)　「RapidMiner BLOG」

Part1 RapidMiner ダウンロード

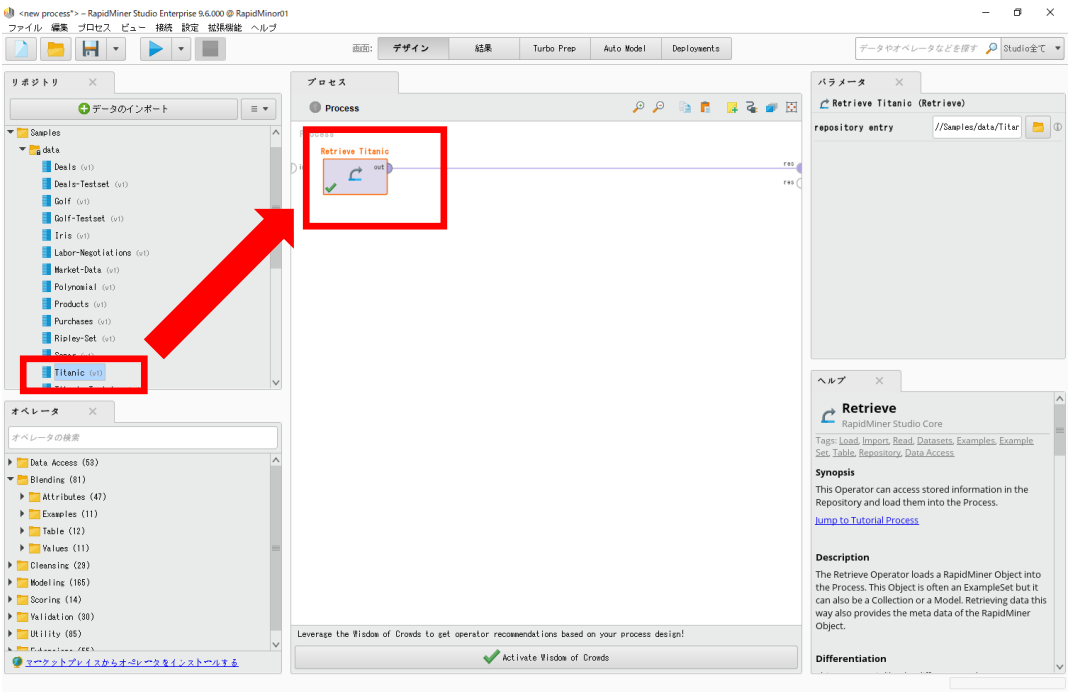
RapidMiner（ラピッドマイナー）は、プログラミングの知識がなくても、
データサイエンティストが行うような高度な分析業務を
ドラッグ&ドロップの簡単な操作で行うことができる、
世界中で使われている機械学習プラットフォームです。

まずは以下を参考に無償版のRapidMinerをダウンロードしてみてください。
https://www.ksk-anl.com/blog/rapidminerの始め方～10stepでできる簡単インストール方法～

Part2 データ読み込み

□ステップ 1/5
RapidMiner へようこそ！
RapidMiner は多くの機能を有しており、ここではRapidMiner Studio を使った基本的な操作方法を説明します。具体的にはデータへのアクセス、データの変換、統計モデルの構築といった事をタイタニック号の乗客データを使って行います。

□ステップ 2/5
データの取り込み
それではタイタニック号の乗船者データを取り込んでみましょう。
ACTIVITY(アクティビティ)
1. 左側にあるリポジトリの欄を参照します
2. Sample フォルダの中のdata フォルダを開きます
3. “Titanic”のデータセットをプロセスエリアにドラッグ＆ドロップします。



□ステップ 3/5
はじめてのプロセス構築
RapidMiner ではオペレータ同士を接続してプロセスを作成します。オペレータ同士はそれぞれのポートを繋げます。
（オペレータ：デザイン画面の左下に表示されている、各アルゴリズム）

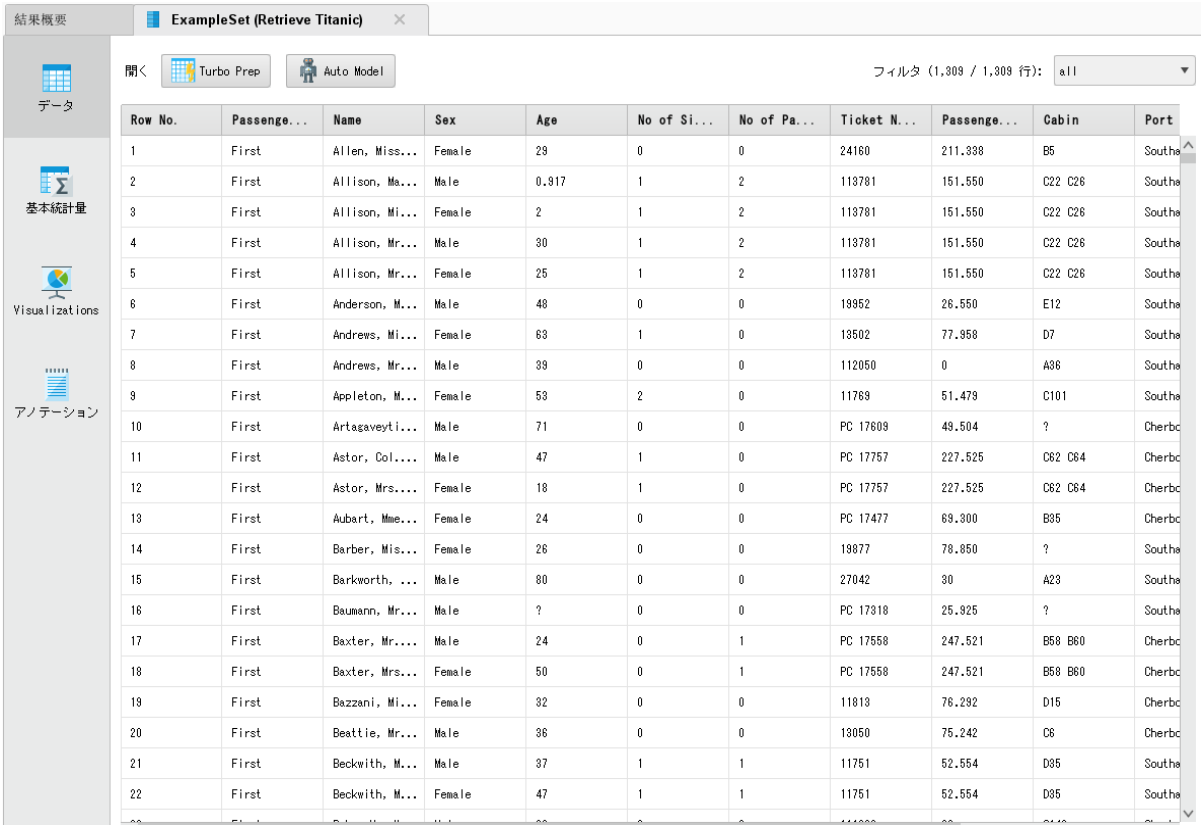
ACTIVITY(アクティビティ)
“Retrieve Titanic”のout ポートとプロセス右側のres ポートとを接続します。out ポートをクリックしてres ポートまで線で繋ぎます。(点線上をなぞるイメージです)

EXPLANATION(説明)
ポートについて

オペレータにくっついている半円のことを「ポート」と呼びます。
情報が出入りする場所です。
例えば‥
out （out）
そのオペレータに入ってきたデータをそのまま出力します。
exa（example）
オペレータで処理したデータセットを出力します。
mod（model）
オペレータで処理したモデルを出力します。
tra（training）
モデル作成用の訓練データを出力します。
per（performance）
モデルの精度を出力します。
ori (original)
加工前のオリジナルデータのことです。
そのオペレータに入ってくる直前のオリジナルデータを出力します。
res（result）
これだけは、オペレータにくっついているものではなく、
画面の右端の最後につなげるものです。
それぞれのオペレータをつないだ、出力結果を出力します。
色々な繋ぎ方をして、複数のresにつないで、一度に実行しても、
複数の結果が出るので大丈夫です！

□ステップ 4/5
プロセスの実行
プロセスの接続が完了したので、実行ボタンを押して結果を出力することが出来ます。
ACTIVITY(アクティビティ)
実行ボタン(画面上部の青い矢印)をクリックして、プロセスを実行することが出来ます。

EXPLANATION(説明)
RapidMiner で最初のプロセスを実行することが出来ました。実行ボタンをクリックすると
オペレータの内容が実行されます。RapidMiner はプロセスを実行すると”result”ポートに
接続されたデータが表示されます。実行後切り替わって表示される結果ビューの中央には
タイタニック号乗船者の家族構成や年齢などのデータがあります。
基本統計量(Statistics) タブをクリックすると要約統計量や有益な情報が表示されます。



Row No.	Passenge...	Name	Sex	Age	No of Si...	No of Pa...	Ticket N...	Passenge...	Cabin	Port
1	First	Allen, Miss...	Female	29	0	0	24160	211.338	B5	Southa
2	First	Allison, Ma...	Male	0.917	1	2	113781	151.550	C22 C26	Southa
3	First	Allison, Mi...	Female	2	1	2	113781	151.550	C22 C26	Southa
4	First	Allison, Mr...	Male	30	1	2	113781	151.550	C22 C26	Southa
5	First	Allison, Mr...	Female	25	1	2	113781	151.550	C22 C26	Southa
6	First	Anderson, M...	Male	48	0	0	19952	26.550	E12	Southa
7	First	Andrews, Mi...	Female	63	1	0	13502	77.958	D7	Southa
8	First	Andrews, Mr...	Male	39	0	0	112050	0	A36	Southa
9	First	Appleton, M...	Female	53	2	0	11769	51.479	C101	Southa
10	First	Artagaveyti...	Male	71	0	0	PC 17609	49.504	?	Cherbo
11	First	Astor, Col....	Male	47	1	0	PC 17757	227.525	C62 C64	Cherbo
12	First	Astor, Mrs....	Female	18	1	0	PC 17757	227.525	C62 C64	Cherbo
13	First	Aubart, Mme...	Female	24	0	0	PC 17477	69.300	B35	Cherbo
14	First	Barber, Mis...	Female	26	0	0	19877	78.850	?	Southa
15	First	Barkworth, ...	Male	80	0	0	27042	30	A23	Southa
16	First	Baumann, Mr...	Male	?	0	0	PC 17318	25.925	?	Southa
17	First	Baxter, Mr....	Male	24	0	1	PC 17558	247.521	B58 B60	Cherbo
18	First	Baxter, Mrs...	Female	50	0	1	PC 17558	247.521	B58 B60	Cherbo
19	First	Bazzani, Mi...	Female	32	0	0	11813	76.292	D15	Cherbo
20	First	Beattie, Mr...	Male	36	0	0	13050	75.242	C6	Cherbo
21	First	Beckwith, M...	Male	37	1	1	11751	52.554	D35	Southa
22	First	Beckwith, M...	Female	47	1	1	11751	52.554	D35	Southa

□ステップ 5/5

ステップのまとめ

おめでとうございます！最初のチュートリアルを完了しました。内容を簡単に振り返ってみましょう。

・オペレータはそれぞれのアクションを実行します。

・オペレータ同士を接続することで、プロセスを構築することが出来ます。

・”res”ポートに接続すると、オペレータの実行結果を見ることができます。

・プロセスを実行するとすべてのオペレータが結果ビューに表示されます。

Challenge(追加質問)

画面左上の「データのインポート(Add Data)」ボタンをクリックすると、

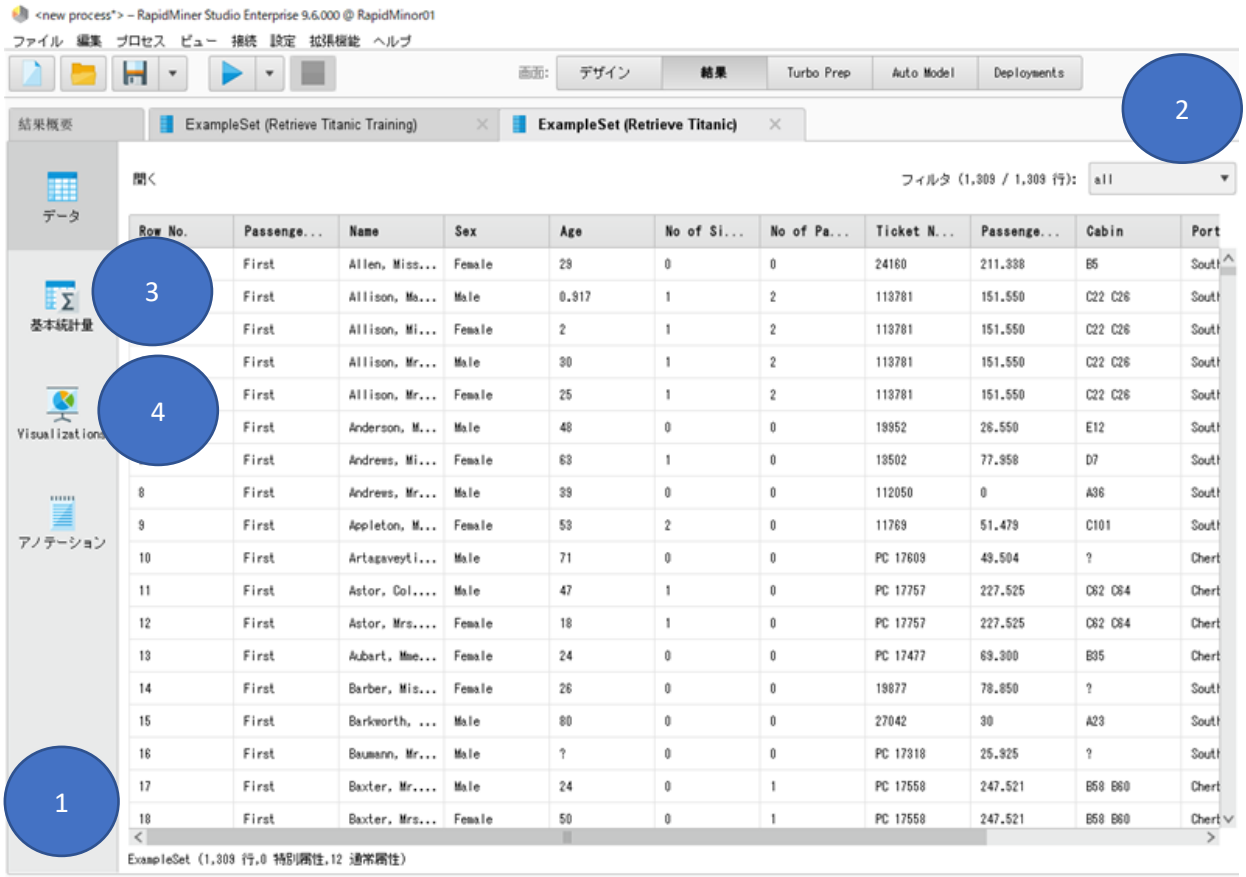
ご自身で準備したデータを読み込むことができます。

Part3 データ可視化

□ステップ 1/1
データ可視化
読み込んだデータをRapidMinerで可視化し、
データを俯瞰的に眺め知見を得る方法を学んでいきましょう！

下の図に示す「結果」画面の機能を順番に確認していきましょう。

1. 結果画面
2. データフィルタ
3. 統計情報
4. グラフ

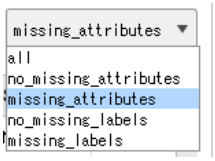


1. 結果画面

結果画面の下部を確認すると、次のことがわかります。
・データセットには、1,309行(example)のデータがあります。
・データセットには、12の項目が含まれます。

2. データフィルタ

データフィルタを使えば、一覧に表示するデータをフィルタリングすることが可能です。
例えば「missing_attributes」を選択すると、項目(attributes)の値が「？」表示の
レコードのみに絞られます。
他の種類のフィルタも試してみましょう。



3. 統計情報(Statistics)

次に、「Statistics」をクリックします。
統計情報では、属性(変数)の型、各属性の欠損値の数、
基本統計量（最小値、最大値、最頻値、平均値、標準偏差）を確認することが出来ます。

EXPLANATION(説明)
データの型について
nominalとnumericの二種類存在します。

〇nominal（テキストや文字列）
polynominal（複数の文字列の値）
red,blue,yellowなど、複数の値を持つ文字列を表します。
binominal（二値）
true⇔false、yes⇔noなどの2つの値を持つ、文字列です。

〇numeric（数値）
　integer（整数）
　　２３、－５、１０２４、７６８など‥
　real（実数）
　　1.23、0.00001など‥
　date_time（時間を含む日付）
　　23/12/2014 17:59など
　date（日付のみ）
　　23/12/2014など
　time（時間のみ）
　　17：59など

「データの型」や「データの種類」を意識して、
自分の分析対象のデータに、
どのような特徴があるのかを掴んでみましょう！

4. グラフ(visualizations)

　次に、「visualizations」をクリックします。
　利用できるグラフがすべて表示されます。
　新しいグラフでデータを可視化してみましょう。

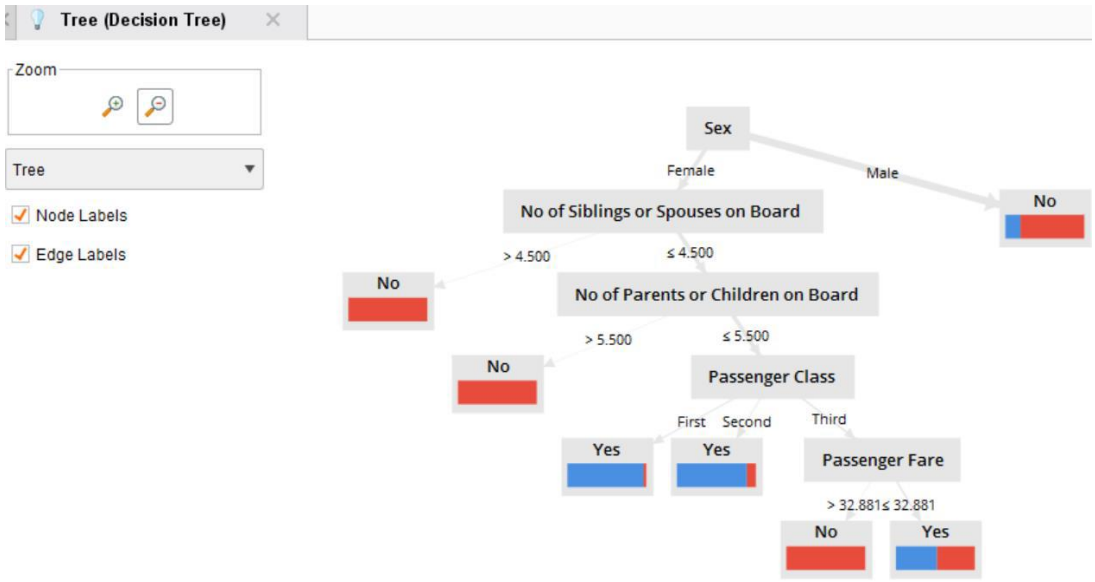
Part4 モデル作成

□ステップ 1/2
生存予測モデルの構築

このチュートリアルでは、最も広く使われている機械学習の一つである
決定木モデル(Decision Tree)を使ってタイタニック号事件の生存者の予測を行います。
もちろん沈没後の今、私たちに出来る事は多くありませんが、同様の状況を
モデルに予測する事はできます。あなたが家族で旅行していたとして本当に３等
客室のチケットを買うべきでしょうか？確かめて見ましょう。

モデル構築後にプロセスを実行すると、以下のような結果が得られます。
性別、客室の等級、家族の人数などの要素で生存率(YES：生存、NO：死亡)が分析できます。

<予測結果>



EXPLANATION(説明)
女性にとっては”family size”が”passenger class”より重要であるということは興味深い事で
す。男性においてはこのパターンは検出されませんでした。女性や子供が優先されるの
で、一般的に男性が生き残る可能性は低くなります。

□ステップ 2/2
ステップ 1/2の予測結果が得られるようにモデル構築を行いましょう。

ACTIVITY(アクティビティ)
以下のヒントを参考にご自身でモデル構築を行ってみてください。

〇ヒント
・“Titanic”のデータセットを使用して分析を行ってください。
“Select Attributes”オペレータを使用し、必要項目のみに絞り込みましょう。

・今回のモデルで予測したい項目(目的変数)は何でしょうか?
’Set Role’オペレータを使用し、定義してみましょう。

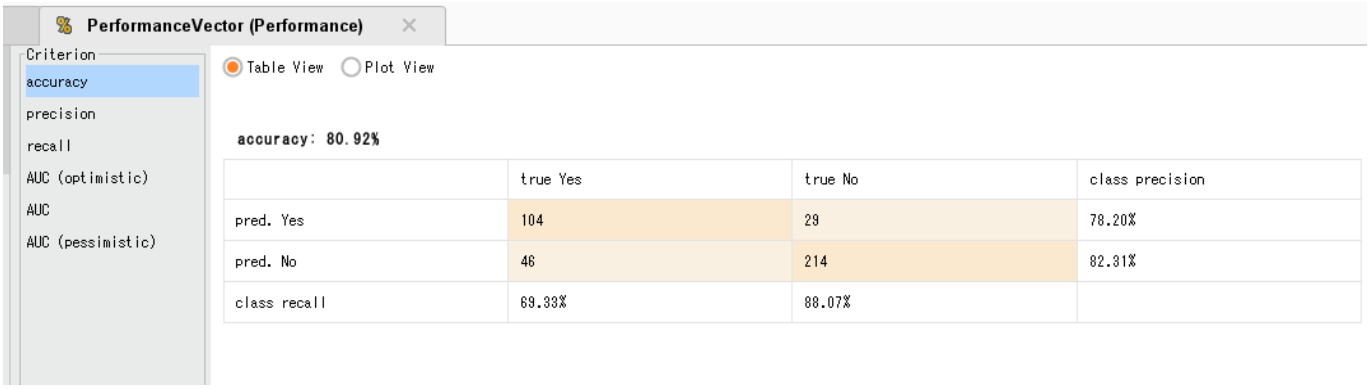
・今回使用する分析手法は決定木モデル(Decision Tree)です。
作成した決定木の深さを制限する、すなわち複雑さを軽減することができます。

Part5 予測モデルの評価

□ステップ 1/2
予測モデルの評価
予測モデルを構築した後に尋ねるべき最も重要な質問は、「このモデルはどのくらいよく機能するのか？」です。将来にかつて出会ったことがないようなシナリオに対して、作成したモデルがよく機能するかどうかをどのようにして述べることができるでしょうか？これを正確に実行する方法はいつも同じです。それはラベル付きのデータのうちいくつかを取っておき、モデル構築に使用しない方法です。このデータはまだラベル付きなので、予測と実際の結果を比較することができます。そしてどのくらいモデルが正確であったかを計算することもできます。このチュートリアルでは、どのようにしてこの検証を実行することができるのかをみていきましょう。

今回は、ラベル付きデータを学習データとテストデータの2 つに分割し、
混同行列（confusion matrix）を使って「Part4 モデル作成」のモデルを評価します。

<評価結果>



	true Yes	true No	class precision
pred. Yes	104	29	78.20%
pred. No	46	214	82.31%
class recall	69.33%	88.07%	

EXPLANATION(説明)
テストデータにおけるモデルのパフォーマンスです。
画面の左側にある”Criterion”で様々なパフォーマンスの測定値を選択することができます。
モデルが全体のどれくらい正確であるかがわかります。混同行列(confusion matrix)は
様々なエラーを示します。例えば、実際は”yes”なのに”no”と予測されたものが29ケースありました。正確度(accuracy)とは、左上と右下という対角線上の数字の和をすべての数の和で割ったものです！この対角線の数字が大きければ大きいほど、モデルのパフォーマンスは良くなります。パフォーマンスを向上させるには、使用するデータの精度が大切です。
欠損値補完や不均衡データの処理などを行ったデータを使用するようにしましょう。

□ステップ 2/2
ステップ 1/2の評価結果が得られるようにプロセスを作成しましょう。

ACTIVITY(アクティビティ)
以下のヒントを参考にご自身でプロセス作成を行ってみてください。

〇ヒント
・「Part4 モデル作成」のプロセスにオペレータを追加してください。

・'Split Data'オペレータを追加し、ラベル付きデータを
学習データとテストデータに分割しましょう。(例えば70/30の比率)

・'Apply Model' オペレータで予測モデルを使用してデータをスコアリングすることができます。

・'Performance'オペレータを追加し、モデルがどれくらいよく機能するのかを計算しましょう。

以上