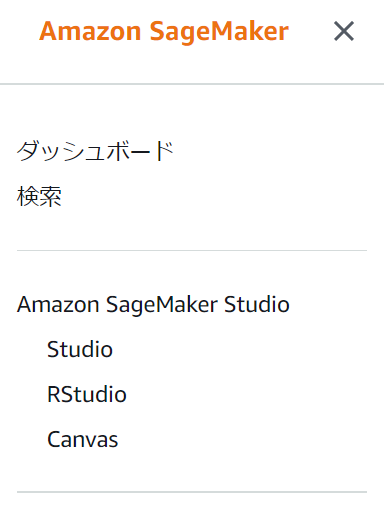
Amazon SageMaker Canvas ワークショップ

2022/02/09

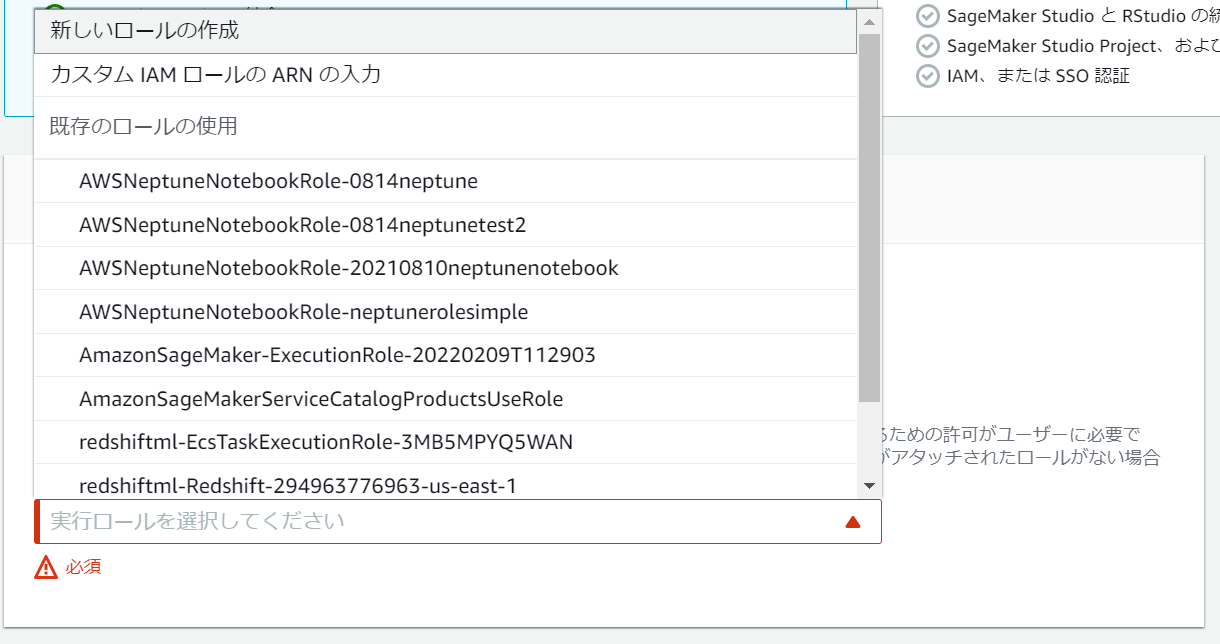
シニアエバンジェリスト 亀田

[はじめに] Amazon SageMaker Canvas はAWSが提供する構造化データを対象とした機械学習をノーコードで実現できるサービスです。このサービスには初回起動から2か月間の無料利用枠が設定されており、本ワークショップシナリオは無料期間内であれば費用は発生しません。ただしSageMakerそのものが用いるストレージの料金は少額が発生しますのでご注意ください。それを過ぎた場合、1回の学習実行で最低**30ドル**の費用が発生することに注意してください。

1. SageMakerのマネージメントコンソールにアクセスします。2022/02/09現在東京リージョンではCanvasに対応していないため、バージニア北部、オハイオ、オレゴンなどを選択します
2. 左ペインからCanvasをクリックします



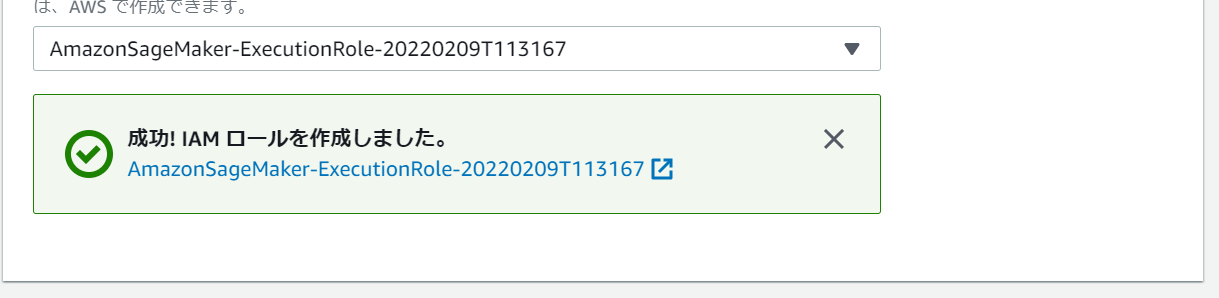
1. ユーザープロファイルから[新しいロールを作成]を選びます



1. [任意のS3バケット]を選びます



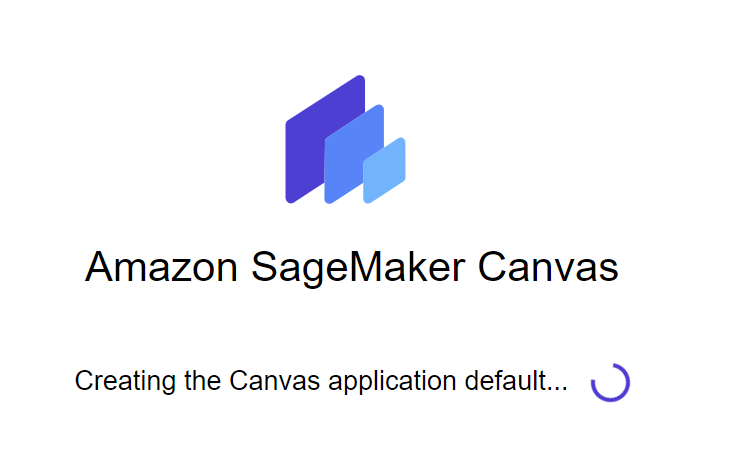
1. [ロールの作成]をおします
2. 以下のように正しくロールが設定されていれば、[送信]をおします



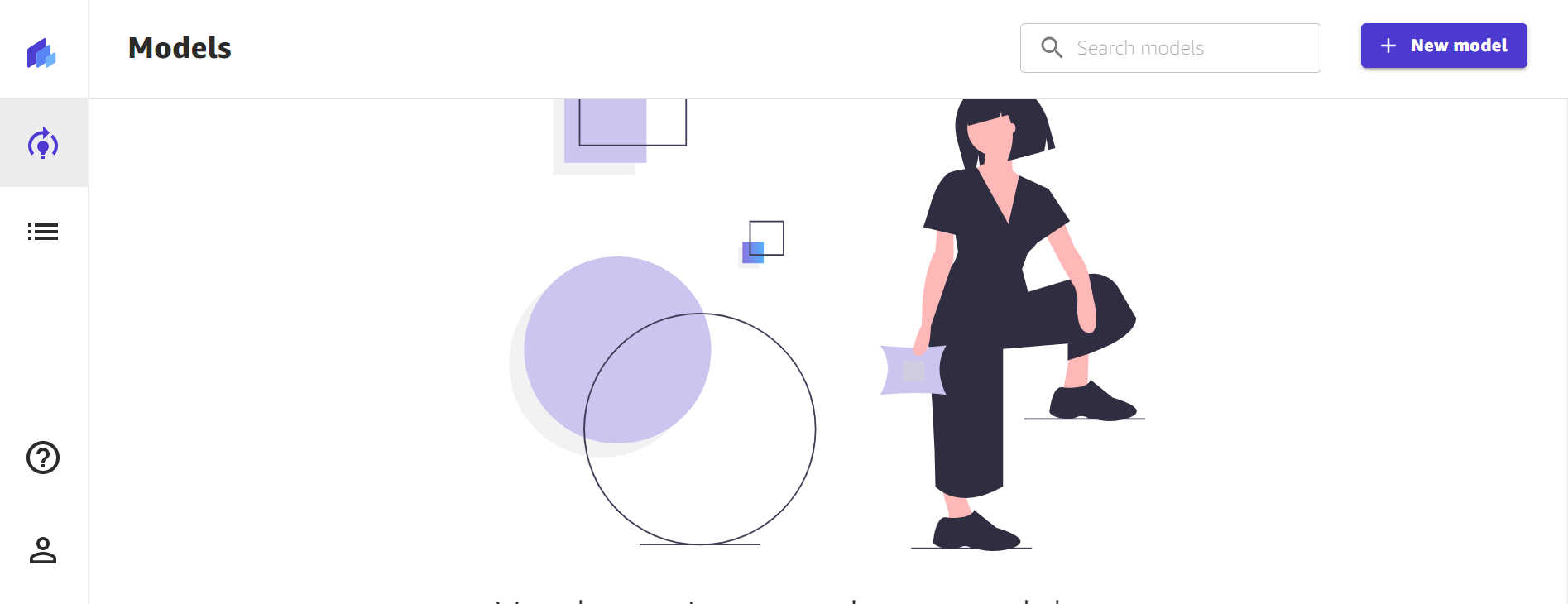
1. SageMaker Domainの起動中になりますので、起動完了まで待ちます。Domainとは開発環境であるCanvasやJupyter Notebook、EFSストレージ、IAMロール、開発環境へアクセスできるユーザーなどが設定としたまとまったSageMakerの設定単位のことを言います。
2. 10分程度でDomainの設定作業が完了しますので、アプリケーションを起動から[Canvas]を選びます



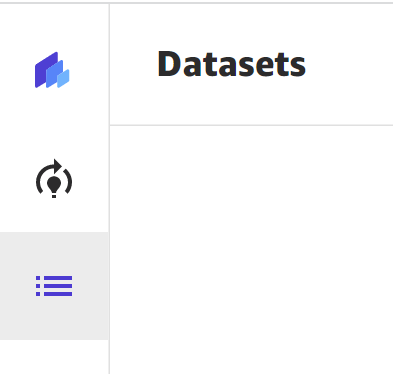
1. 以下のようにCanvasの起動画面が表示されますので、起動が完了するまで待ちます。



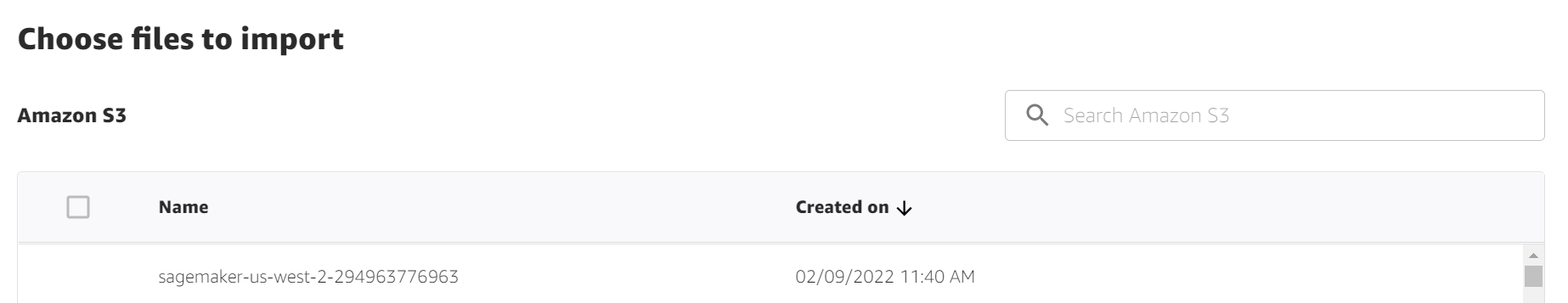
1. 以下の画面が表示されれば起動完了です。



1. 左ペインから[Datasets]をクリックします



1. 画面右上[Import]ボタンをおします
2. 以下のようにCanvas起動時に自動でS3バケットが作成されていることがわかります。ブラウザの別タブでS3マネージメントコンソールにアクセスします



1. バケット名を特定しクリックします
2. Githubから[revised\_boston\_house\_prices.csv]をダウンロードします。このデータはscikit-learnで公開されている米国ボストンの住宅価格を予測するために用いられるデータです。オリジナルデータは、一部差別につながる可能性のある不適切なデータを含んでいるため、その部分を削除しています。

それぞれのカラムは以下を意味しています。

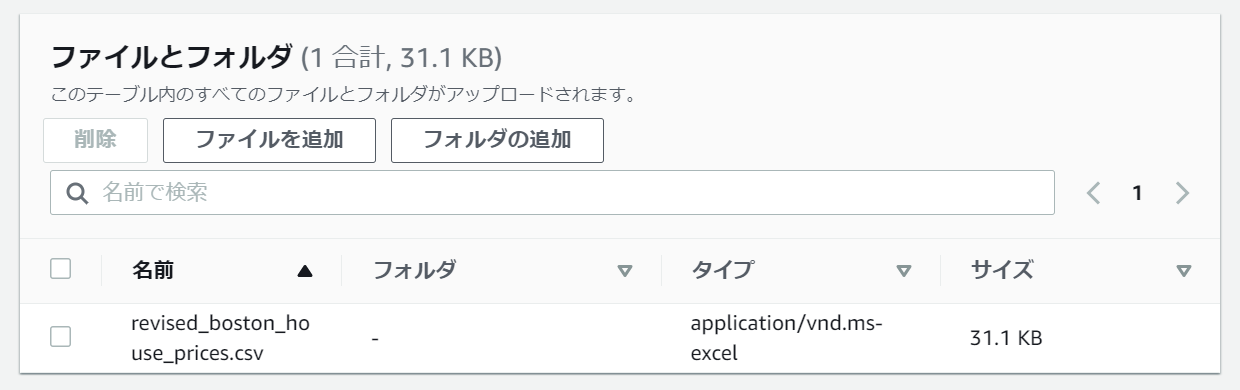
|  |  |
| --- | --- |
| **CRIM** | 人口 1 人当たりの犯罪発生数 |
| **ZN** | 25,000 平方フィート以上の住居区画の占める割合 |
| **INDUS** | 小売業以外の商業が占める面積の割合 |
| **CHAS** | チャールズ川によるダミー変数 (1: 川の周辺, 0: それ以外) |
| **NOX** | NOx の濃度 |
| **RM** | 住居の平均部屋数 |
| **AGE** | 1940 年より前に建てられた物件の割合 |
| **DIS** | 5 つのボストン市の雇用施設からの距離 (重み付け済) |
| **RAD** | 環状高速道路へのアクセスしやすさ |
| **TAX** | $10,000 ドルあたりの不動産税率の総計 |
| **PTRATIO** | 町毎の児童と教師の比率 |
| **LSTAT** | 給与の低い職業に従事する人口の割合 (%) |

これらのパラメーターを変更した場合、リアルタイムで住宅価格の平均値（MEDV）を予測するモデルを作ることが今日のハンズオンのゴールです。

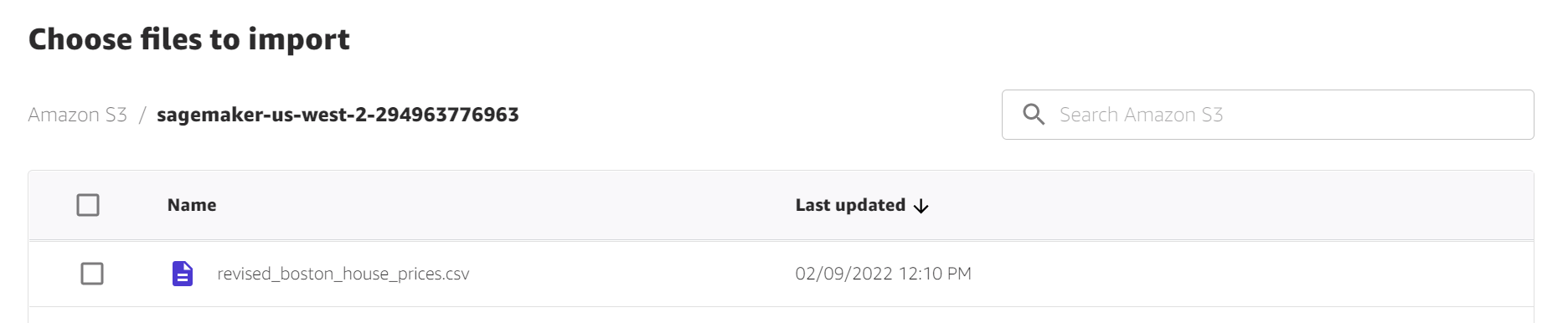
1. S3バケット詳細画面より[アップロードボタン]をおします

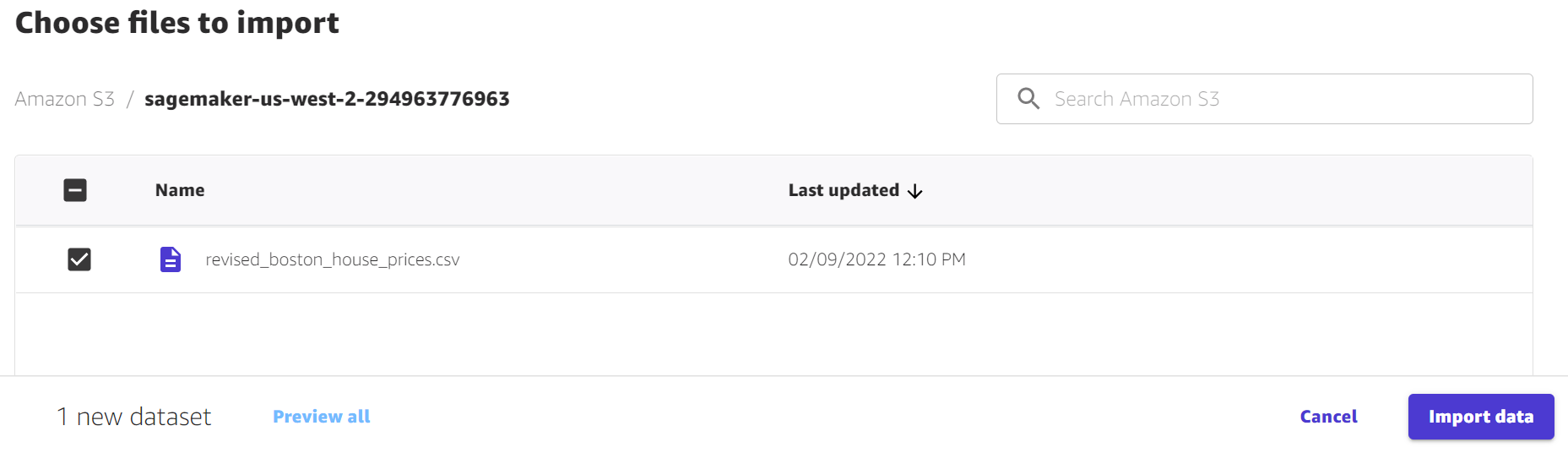


1. ファイルの追加、より先程ダウンロードしたCSVをアップロードします

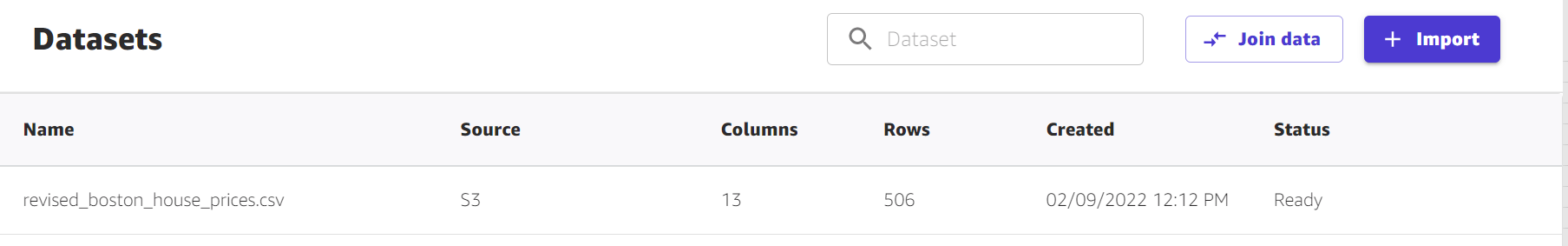


1. 設定などは全てデフォルトのまま[アップロード]ボタンをおします
2. ブラウザ別タブですでに開いているCanvasの画面に戻り、S3バケットをクリックします
3. 以下のようにCSVが認識できていれば、チェックをつけて[Import data]をおします

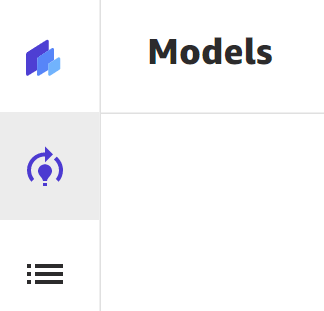




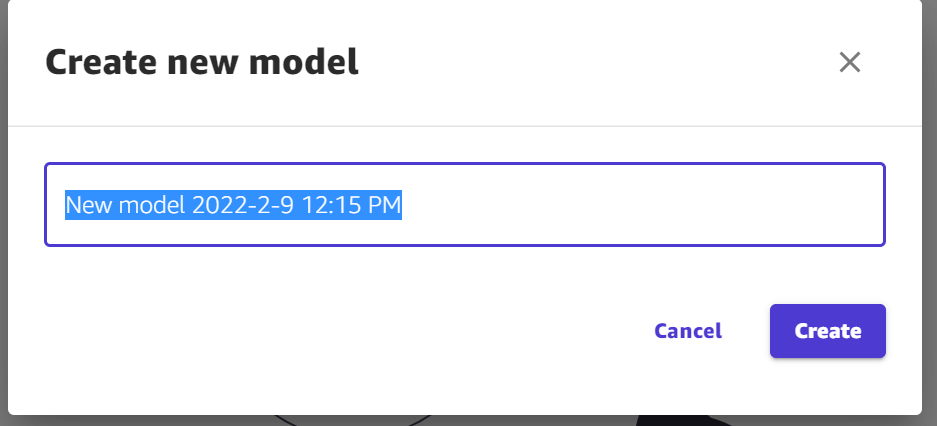
1. しばらく待つとデータ取り込みが完了し、Statusが[Ready]になれば完了です



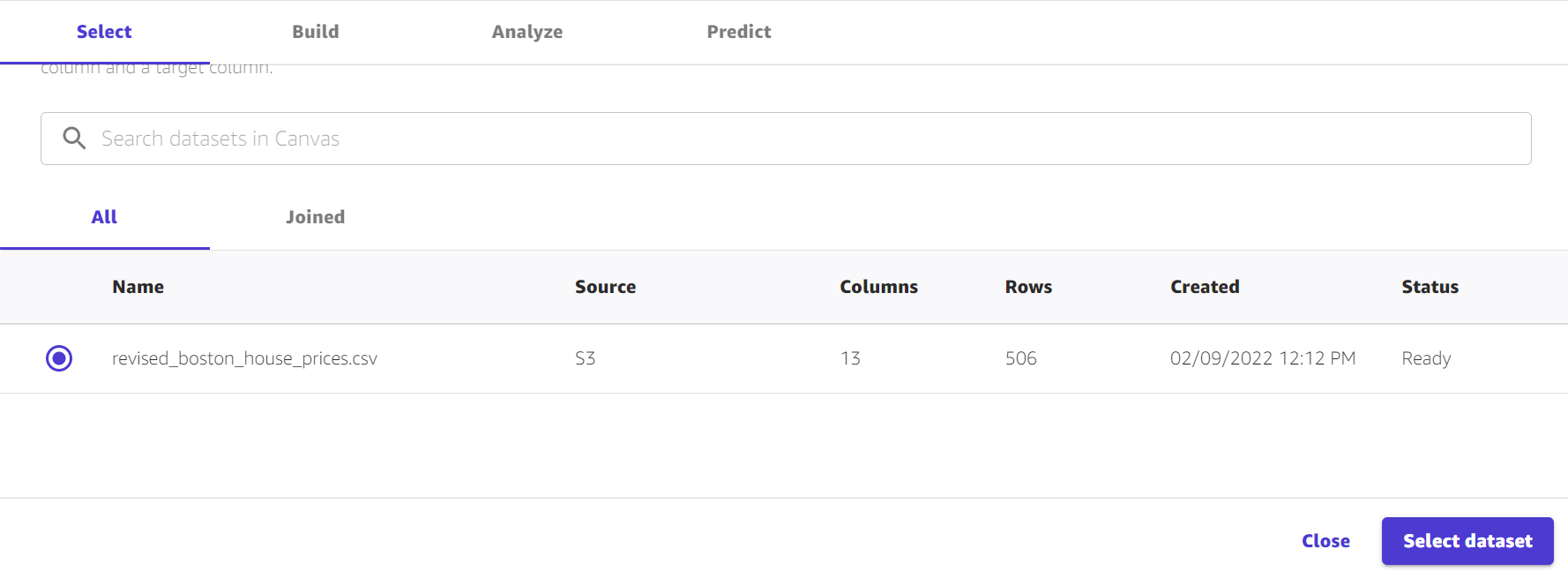
1. 次に左ペインから[Modelをクリックします



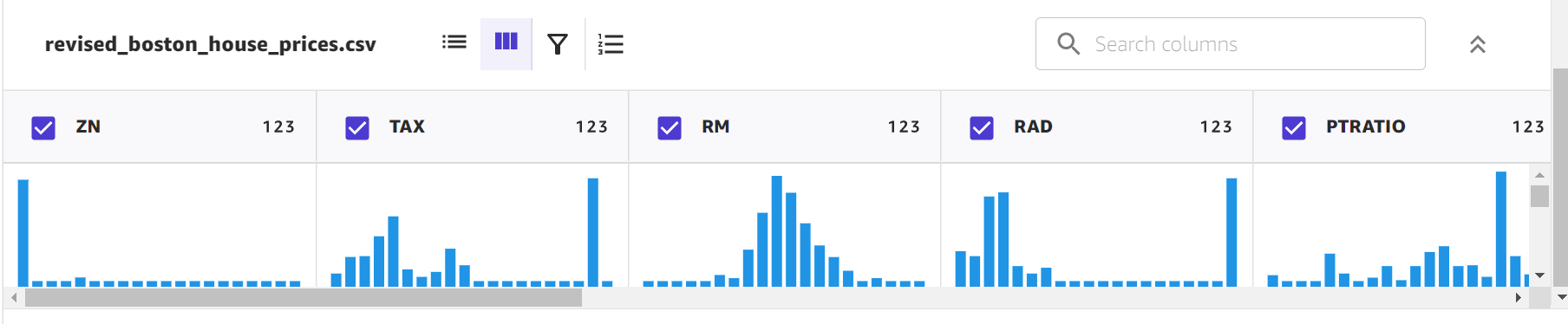
1. 画面右上[+ New model]をおします
2. 自動で名前が、時間をベースについていますのでそのまま[Create]をおします



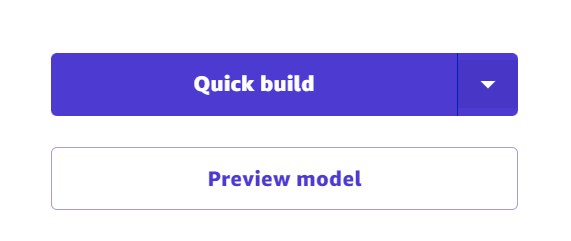
1. Intro ダイアログが表示されますが。Skipします
2. [Select]タブから、先程設定したcsvファイルを選択し、[Select dataset]をおします
3. **Select a column to predict**



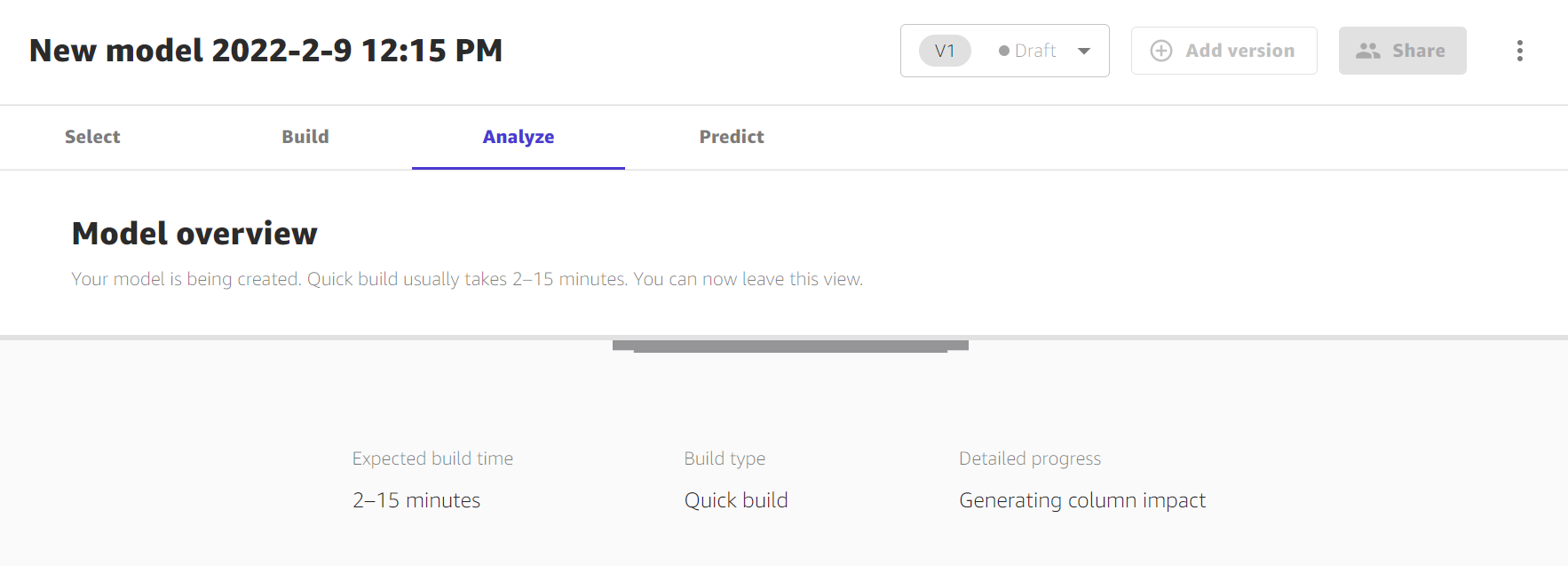
1. [Build]タブに移動し[Select a column to predict]から[MEDV(住宅価格平均)]を指定します。
2. [Model type]は自動で重回帰分析であるNumericが指定されていますが、[Change Type]を指定すると、2値分類、多値分類、時系列予測がサポートされていることがっ分かります。なお、時系列予測はAmazon ForecastをCanvas経由で起動する形態となっています。
3. 画面下部分では、学習に用いられるデータが偏りを含んでいないか等データ分布状態を確認できるようになっています。



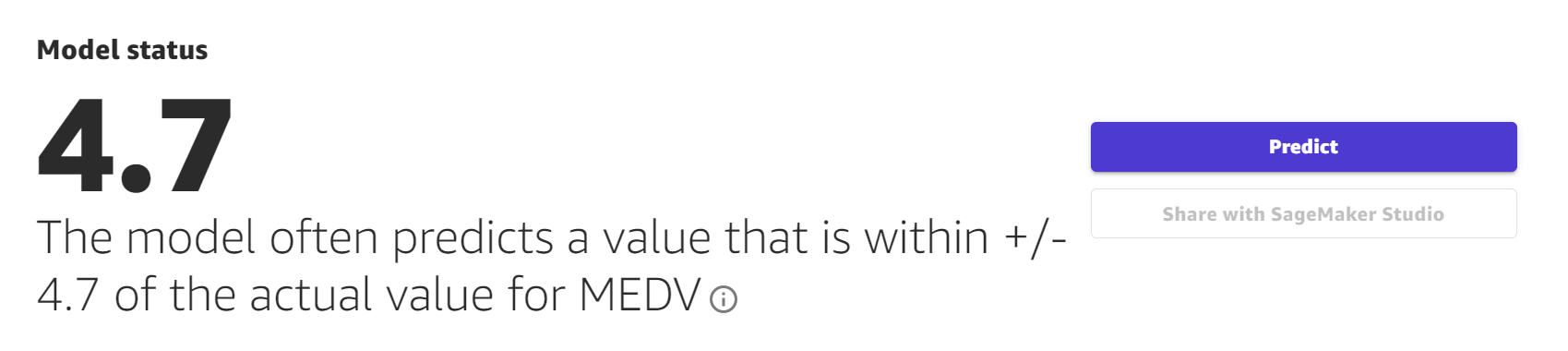
1. [Quick Build]をおします



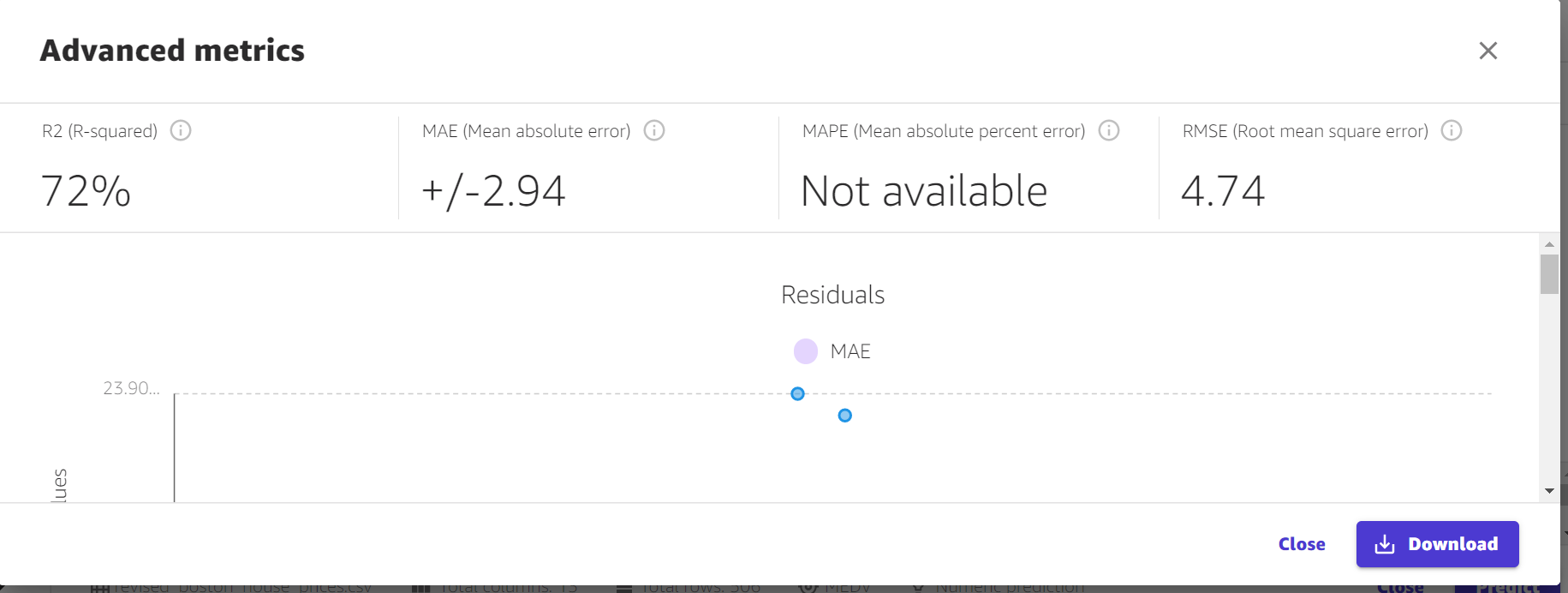
1. 学習が開始されますので、15分程度まちます。なおこのハンズオンでは、処理速度を優先させるためQuick Buildを選んでいますが、商用環境での利用を検討される場合より精度が高くなるStandard Buildの指定をおすすめします



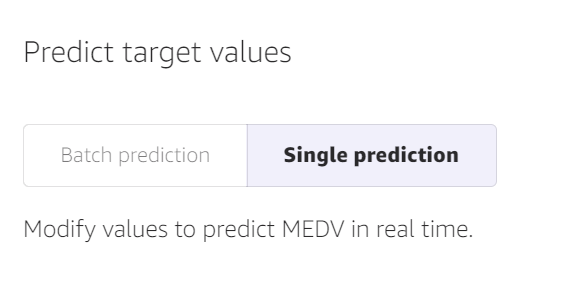
1. 学習が完了すると、以下の通りモデル指標が表示されます。この値はRMSE(Root Mean Squared Error)と言われるもので、回帰分析においては数字が小さければ小さいほど精度が高いとされる指標です。このハンズオンでは500+件のデータをベースにQuick Buildでモデルを作成しているため、それほど精度が出ていないことに注意してください。



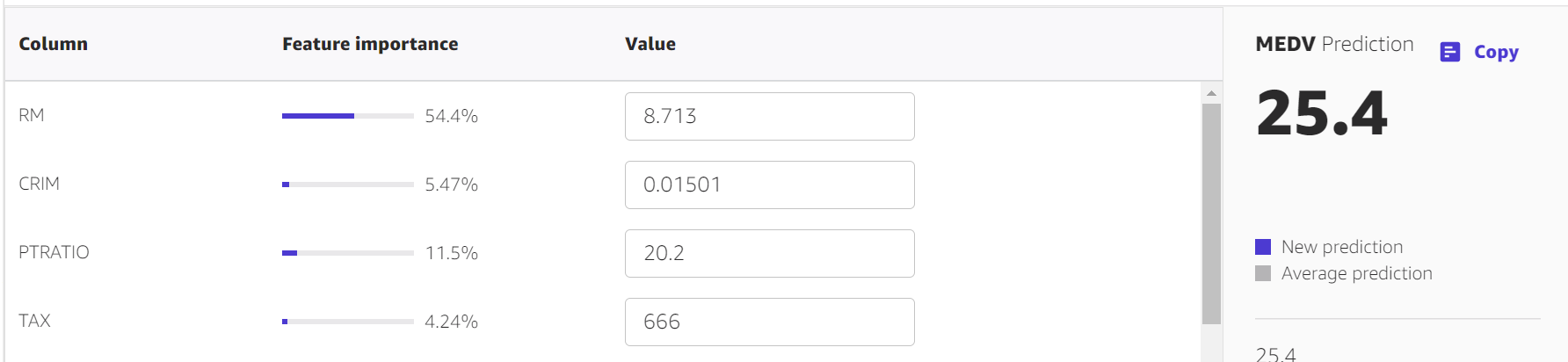
[Advanced Metrics]からより詳細のモデル分析情報を入手可能です。



1. [Predict]ボタンをおします
2. [Single Prediction]をクリックします

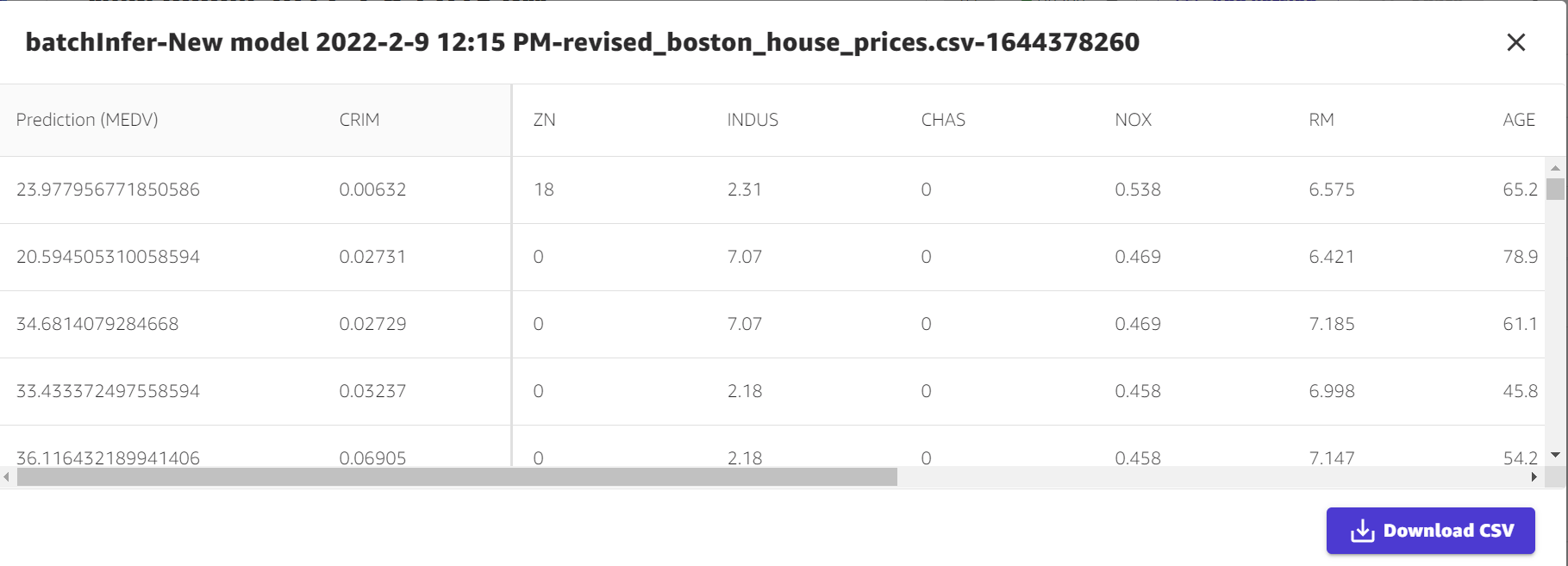


1. [Feature importance]とは、予測に影響を与えている各特徴量毎の割合を示します。より高い数値のカラムが、より強く予測に影響を与えています。それぞれの値を変更して、画面右[Update]ボタンをおすとMEDVの値が推論されます。



1. (Option)[Batch Prediction]では、複数の予測したい値をcsvなどで準備しアップロードすることで、バッチ予測（複数の値をまとめて予測）することが可能です。例えば、今日の学習に使用したcsvをそのまま予測用に指定し実行してみてください。

以下のように予測の一覧が表示され、CSVでの出力が可能です。



おつかれさました！

削除は以下の順番でおこなってください

・SageMaker Canvasマネージメントコンソールからユーザーをクリック

アプリケーション削除

・SageMaker Canvasマネージメントコンソールから Domain を削除

S3バケットを削除