# レポート概要

## 1. 背景

観客数予測は売上最大化に直結する重要な分析です。特にビールはプロ野球観戦の最も重要なおともであり、その在庫管理は売り上げ増加のために非常に重要です。Plus1 Oneのサイトによりますと、概算にはなりますが、一人当たり2杯のビールを試合の度に購入しているとのデータもあります。正確な観客数予測を行うことで、在庫管理を最適化し、売上の最大化を図ります。本レポートでは、プロ野球の観客数予測とそれを基にした売上予測を通じて、ビールの在庫管理と戦略的な意思決定の最適化を目指します。

## 2. 目的

観客数予測を基にしたビールの在庫管理の最適化と、売上最大化を目指すことが本プロジェクトの目的です。観客数予測の精度を高め、その情報を元に、在庫管理の改善提案を行います。

## 3. データ収集

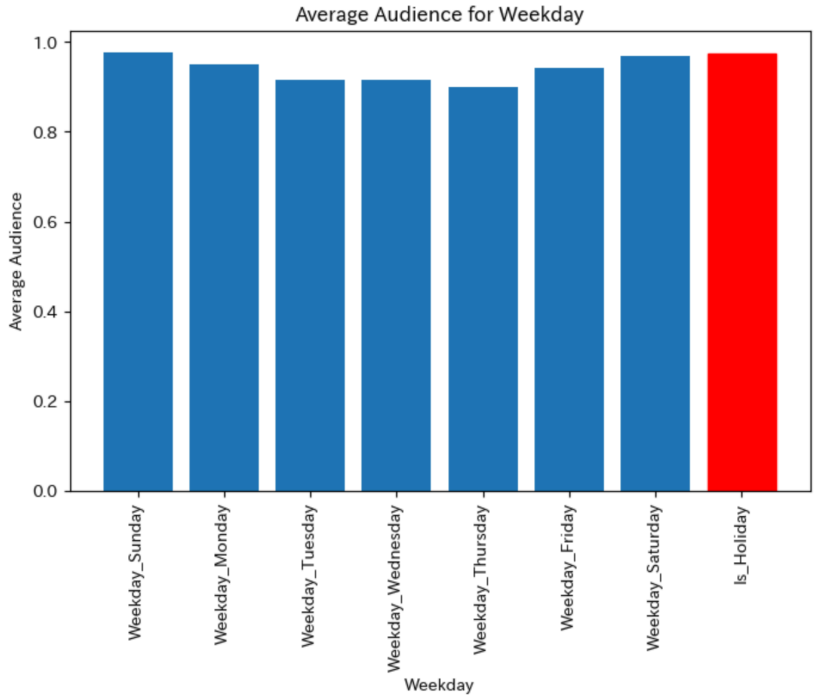
- 観客数: プロ野球の観客数データは、プロ野球Freakというサイトから収集しました。期間は2015年から2024年までのシーズンデータです。
- 天気情報: 気象庁からの福岡市の天気情報を使用し、試合当日の平均気温、平均風速、一日の総雨量を特徴量として取り入れました。期間は2015年から2024年までです。
- ビールデータ: ビールの売上情報は、Plus1 Oneのサイトから獲得したデータを基にしています。単価や仕入れ値などをこのデータから概算しました。

## 4.データの前処理

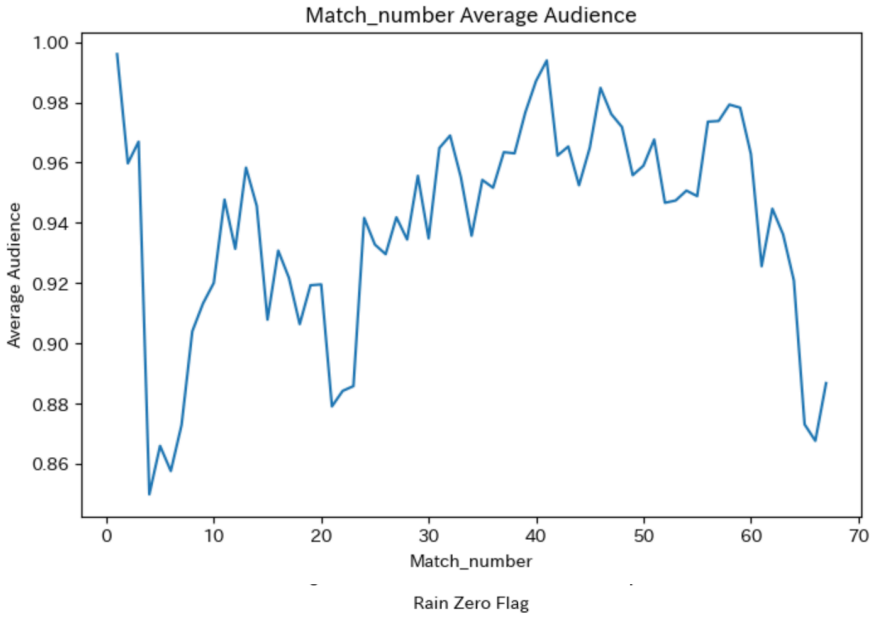
1. 中止試合の排除
   試合結果が「中止」と記録されているデータは、観客数や売上に影響を与えないため、排除しました。
2. 該当ドームのみの抽出
   観客数予測の対象となるのは、特定のドーム（ヤフオクドーム、PayPayドーム、みずほPayPay）で開催された試合のみです。それ以外のデータを排除しました。
3. 日付処理
   試合日の情報を`Date`列として`datetime`型に変換し、曜日の情報も抽出して新たに`Weekday`列を作成しました。
4. 観客数の数値化
   観客数を数値型に変換しました。文字列や欠損値が含まれている場合も、適切に処理しました。
5. 降水量、気温、風速の数値化
   天気データ（降水量、気温、風速）も数値型に変換しました。
6. 不要な列の削除
   重複行である日付データ（`FormattedDate`, `yyyy/mm/dd`）と試合開始前に知ることがない情報（'score'など）、会場に関するデータ（'Venue'）を削除しました。

# 5. 特徴量エンジニアリング

1. 祝日フラグの作成
   試合日が祝日かどうかを示すフラグ(`Is_Holiday`)を作成しました。日本の祝日判定ライブラリを使用しています。




2. 試合数の作成
   同年内で何試合目かを表す特徴量(`Match_Number`)を作成しました。これは年ごとに累積カウントすることによって実現しました。
3. 雨量**0**のフラグ
   降水量が0mmの場合に`Rain_Zero_Flag`としてフラグを立てました。これによる占有率



（後述）の箱ひげ図は以下。

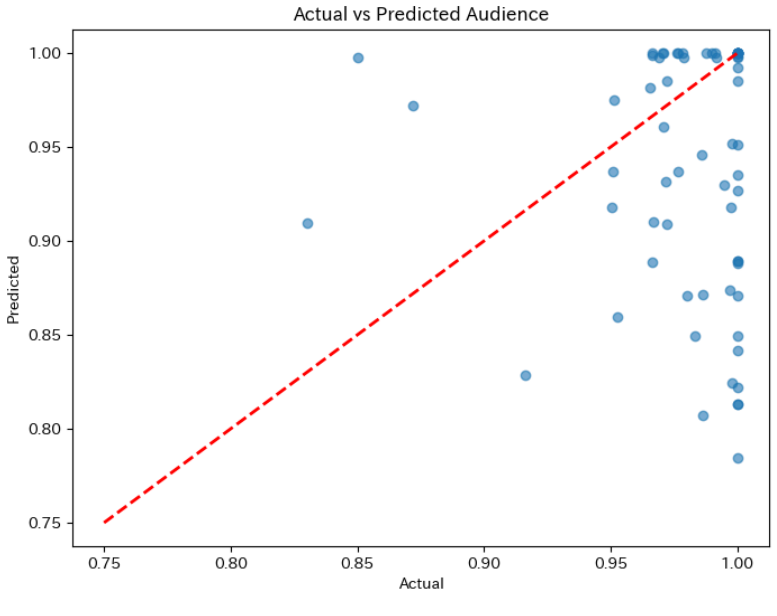
4. ドームの定員
   各ドームの定員情報を新たな特徴量として追加しました。`Year`ごとの定員を定義し、元のデータとマージしました。
5. 占有率の作成
   観客数をドームの定員で割ることによって、占有率（`Occupancy`）を計算しました。また、定員が年によって変化するため、この占有率を目的変数とします。これは、2019年前後で大きく定員数が増加したことを受けてのものです。またこれにより、今後定員数が大きく増減した場合でも対応しやすくしています。

## 6. モデル構築

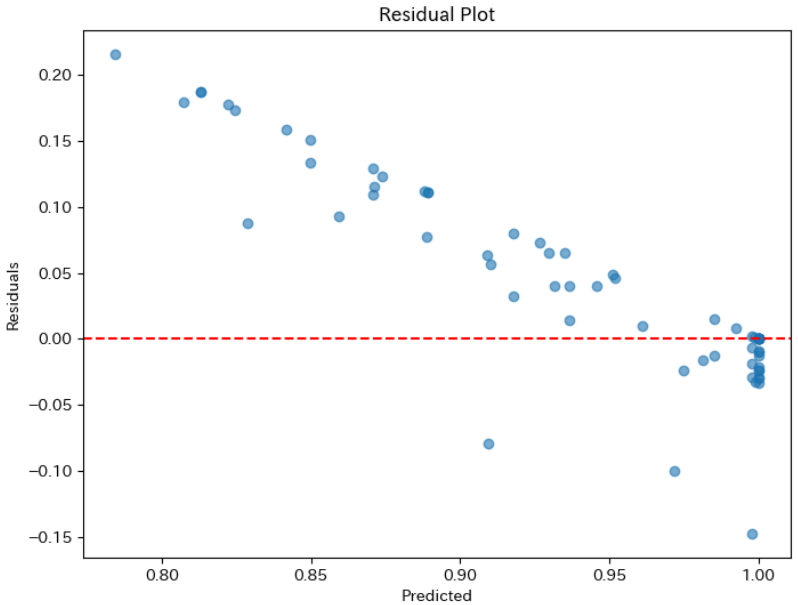
- 使用した特徴量
  - 平均気温、風速、雨量
  - 対戦チーム、曜日、祝日、雨の有無
  - 何試合目か
- 学習データとテストデータ
  - 学習データ：2015~2023年シーズンデータ
  - テストデータ：2024年シーズンデータ
- 使用した予測モデル
  決定木を使用しました。決定木は解釈性が高く、特徴量間の関係を視覚的に理解しやすいため選択しました。
- モデル評価方法
  RMSE（Root Mean Squared Error）を使用しました。予測精度の評価として広く使用されており、モデルの予測誤差の大きさを測定するのに適しています。
- 評価結果
  RMSEは0.085程度であり、十分な精度を達成した

## 7. 結果の表示（機械学習）

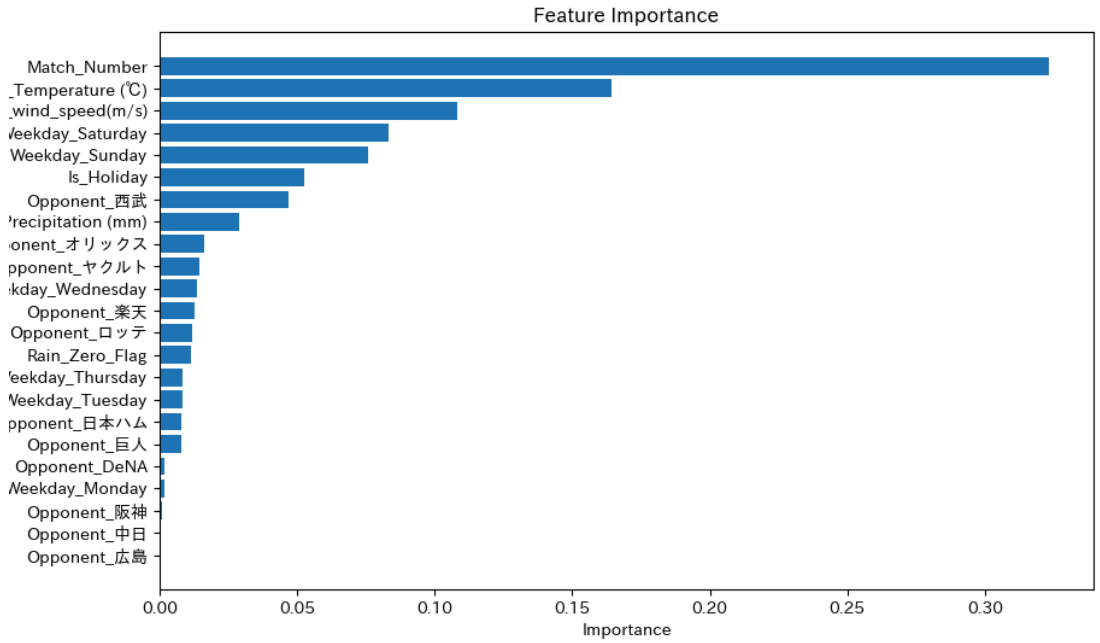
- 実測値と予測値のプロット
  赤線は予測値と観測値が一致する場合の直線である。
  予測結果は赤線の下に多く分布しており、予測は本来よりも低く見積もっている。



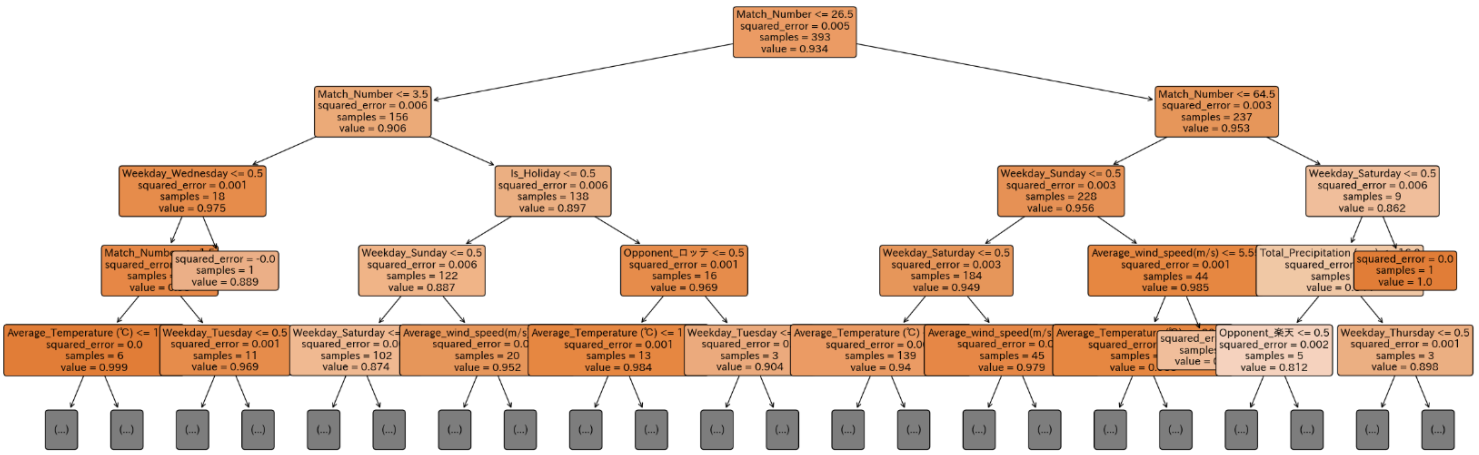
- 残差プロット
  予測値と残差（観測値と予測値の差）をプロットすることで適合性をチェックします。
  予測値が小さくなるほど残差が大きくなり、ランダム性が観測できないので、小さい予測値を持つ列は改善の余地がある。
- 重要な説明変数
  決定木においてサンプルを分けるのに重要となる説明変数数を可視化しました。特徴量エンジニアリングで作成した何試合目かを表すMatch_Numberの重要度が大きいことが分かります。また、気温や風速、曜日に関するデータの重要度は高く、チームに関するデー



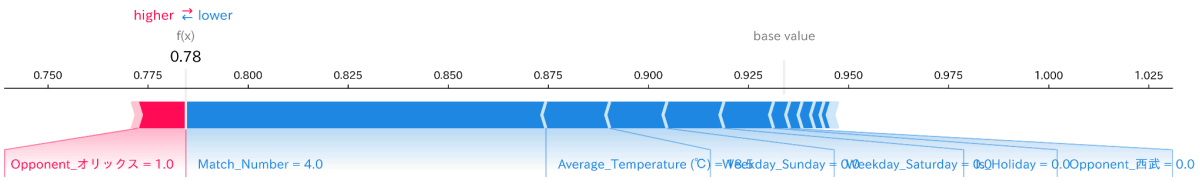
タの重要度は低いです。

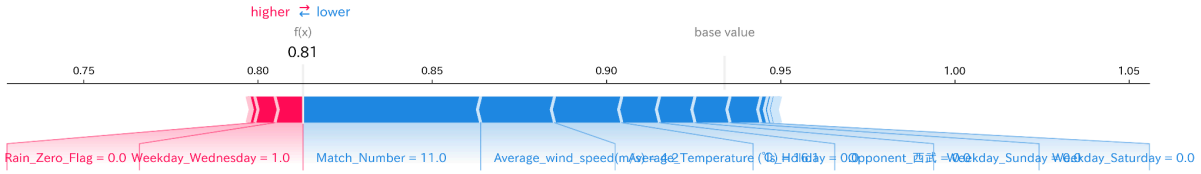


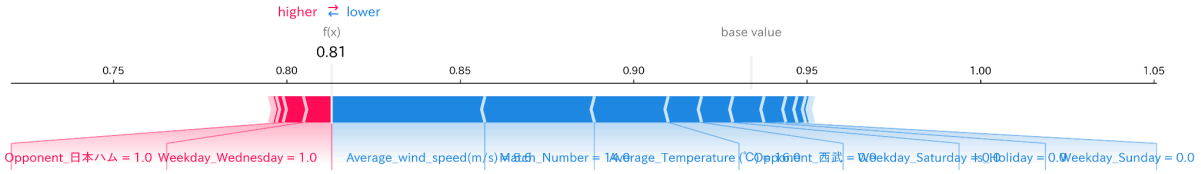

- 決定木の構成（深さ**4**まで）
  深さ4までの決定木の分類の方法を可視化します。
  重要度の高いものは浅い部分の分類分けによく使われています。




- 誤差の大きいサンプルの**SHAP**
  誤差の大きかった3データについて、どのようにその値を取ったかを可視化します。重要度の大きいMatch_Numberが小さい値へ向かわせており、もう少し丁寧な特徴量の作成が必要です。

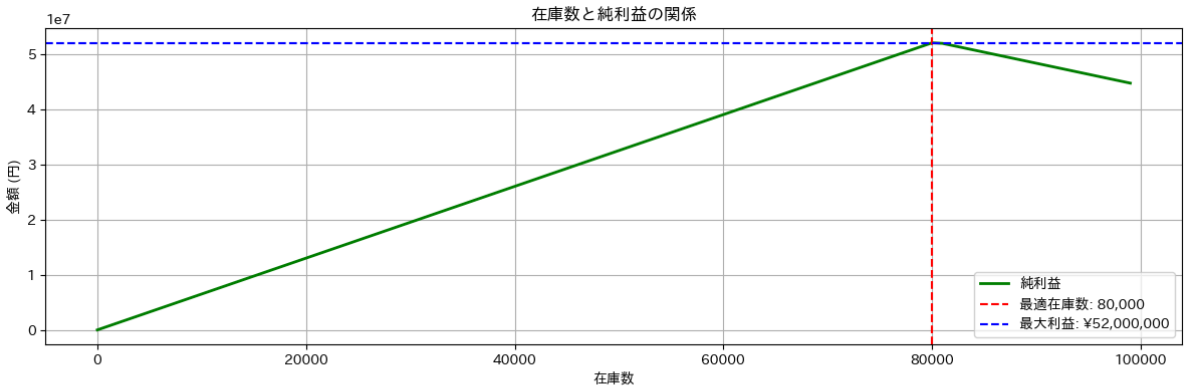


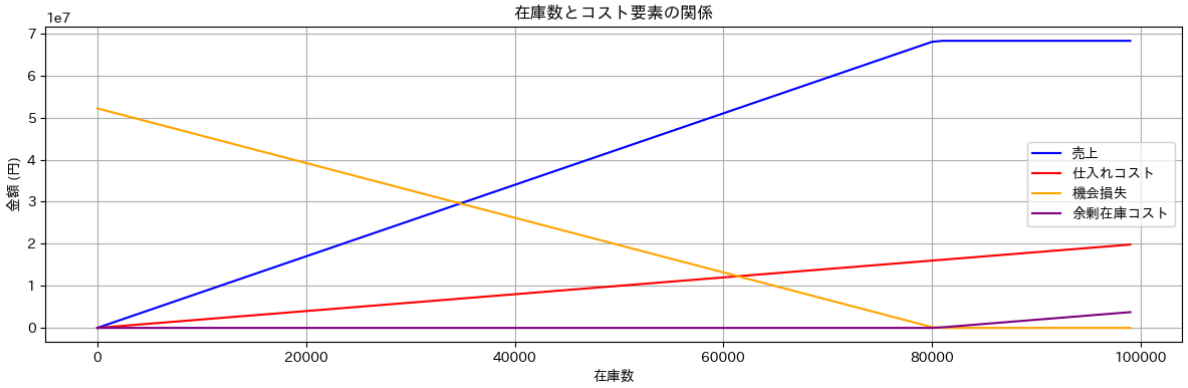




# 8. 結果の提示（ビジネス）

- **1**人当たりの購入回数、単価、仕入れ値
  売上計算に使用しているデータは、Plus1 Oneのサイトからの概算に基づいています。1人当たりの購入回数は2回、ビール単価は850円、仕入れ値は200円と設定しました。
- ビジネス戦略の提案
  観客数予測を基にした在庫管理の徹底を提案しています。予測に基づき、適切な在庫を確保することで、機会損失や過剰在庫を減らし、利益を最大化することができます。
- 視覚化
  観客数予測と売上推定の結果をグラフで視覚化しました。具体的には、機会損失と過剰在庫を折れ線グラフで表現しています。
  このグラフではMatch_Numberが0、つまり1試合目のデータに関して予測したものです。
  予測値が正確に予測できて、さらにこの戦略に従った場合、5200万円もの利益を上げることができます。




# 9. 考察

- 最も重要な課題
  観客数予測において最も重要な課題は、データの不足です。特に現在の在庫数や実際の仕入れ値、一人当たりの購入回数などが明確でないため、これらの情報が不足しています。
- 予測精度の向上
  Match_Numberは高い重要度を持つものの、誤差を大きくする原因でもあるので、 より詳細に設計する必要があります。観客数予測精度を向上させるためには、より多くのデータを収集し、他の特徴量（例：プロモーション効果）をモデルに組み込むことが重要です。

# 10. 展望

- 今後の展開
  他の予測モデル（例：ランダムフォレストやXGBoost）を試し、精度向上を目指します。また、特徴量の改善や追加によって、予測精度をさらに高めることができると考えています。
- 新たな取り組み
  今後は、価格設定やプロモーション効果などをモデルに組み込むことで、売上最大化に直結する意思決定をサポートすることを目指します。
  また、毎日の「リアルタイム予測」や「在庫調整の自動化」を将来的に行うことも可能です。