バイクシェアリング需要予測に関する研究レポート

2025 年 11 月 26 日

1 序論

1.1 テーマの選定理由 (Why you choose the prediction theme)

近年、都市部における交通渋滞の緩和や環境負荷の低減、健康増進を目的として、自転車シェアリングシステム（Bike Sharing System）が世界中で急速に普及している。このシステムは、ユーザーが任意のステーションで自転車を借り、別のステーションで返却できる利便性を提供している。しかし、運営側にとっては、各ステーションにおける自転車の需要を正確に予測し、適切な再配置（リバランス）を行うことが重要な課題となっている。需要過多による機会損失や、需要過少による在庫過多を防ぐためには、気象条件や日時などのデータに基づいた高精度な需要予測モデルが不可欠である。また、利用者にとっても、需要予測によって「この時間は混雑していて借りられないかもしれない」といった情報を事前に得ることができれば、移動計画を立てやすくなるなど、日常生活における利便性向上に直結する。本研究では、機械学習を用いてこの需要予測問題に取り組み、線形モデルと非線形モデルの比較を通じて、最適な予測手法を検討するために本テーマを選定した [1]。

1.2 予測の貢献 (How the prediction contribute)

本予測モデルによって自転車のレンタル需要を正確に予測できれば、以下のような貢献が期待できる。

- **運営効率化**: 需要の高いエリアや時間帯を事前に把握することで、トラックによる自転車の再配置を効率的に行い、運用コストを削減できる。
- **ユーザー満足度の向上**: 「借りたい時に自転車がない」「返したい時にポートが満車で返せない」といった事態を防ぎ、サービスの信頼性と利便性を向上させる。
- **都市計画への応用**: 自転車利用のパターンを分析することで、新たな自転車レーンの整備やステーションの設置場所の最適化など、データに基づいた都市交通計画の策定に寄与する。

2 データ

2.1 データの収集方法 (How to collect data)

本実験で使用するデータセットは、Kaggle のコンペティション「Bike Sharing Demand」から取得されたものである。このデータセットは、米国ワシントン D.C. の Capital Bikeshare プログラムによって提供された、2011 年から 2012 年までの 2 年間の利用履歴データと、対応する気象データで構成されている。

2.2 データセットの概要

データセットには、1 時間ごとのレンタル数（‘count‘）に加え、以下の特徴量が含まれている。

- **日時情報**: 日付（datetime）、季節（season）、休日（holiday）、勤務日（workingday）
- **気象情報**: 天気（weather）、気温（temp）、体感気温（atemp）、湿度（humidity）、風速（windspeed）

提供されたデータセットは、毎月 1 日から 19 日までが学習用（Train）、20 日から月末までがテスト用（Test）として分割されている。しかし、本実験ではテスト用データの正解ラベルが公開されていないため、提供された学習用データ（全 10,886 件）をさらに時系列順に学習用（60%）、検証用（20%）、テスト用（20%）に再分割して使用した。

3 手法

3.1 データ前処理と特徴量エンジニアリング

生データに対して、モデルが学習しやすい形式に変換するための前処理を行った。

3.1.1 対数変換

ターゲット変数であるレンタル数（y）は、0 以上の値をとり、右に長い裾を引く分布（ポアソン分布に近い形状）をしている。回帰モデルの多くは誤差が正規分布に従うことを仮定しているため、ターゲット変数に対して以下の対数変換を適用した。

$$y' = \log(1 + y) \tag{1}$$

これにより、分布を正規分布に近づけるとともに、評価指標である RMSLE の最小化を、変換後の値に対する MSE（平均二乗誤差）の最小化問題として扱うことが可能となる。

3.1.2 特徴量の抽出と変換

- **日時特徴量**: ‘datetime‘ カラムから、年（year）、月（month）、日（day）、曜日（weekday）、時間（hour）を抽出した。特に「時間」は需要に与える影響が極めて大きいため重要な特徴量である。
- **カテゴリカル変数の One-Hot Encoding**: 季節（season）、天気（weather）に加え、数値として表現されているが実質的にカテゴリカルな性質を持つ**月（month）、時間（hour）、曜日（weekday）**に対しても One-Hot Encoding を適用した。例えば、時間は「0」から「23」の数値であるが、「23 時」と「0 時」は時間的に隣接しているにもかかわらず数値的な距離が遠い。また、朝の通勤ラッシュ（8 時頃）と夕方の帰宅ラッシュ（17 時頃）にピークを持つ非線形な構造がある。これを線形モデルで捉えるためには、各時間を独立したカテゴリとして扱うことが有効である。
- **スケーリング**: 気温、体感気温、湿度、風速などの連続値変数は、StandardScaler を用いて平均 0、分散 1 に正規化した。これは、勾配降下法を用いる MLP などの学習を安定させるために重要である。

最終的にモデルに入力される特徴量は以下の通りである。

- **数値変数 (5 次元)**: temp, atemp, humidity, windspeed, year
- **カテゴリカル変数 (One-Hot Encoded)**:
  - season (4 次元): 春, 夏, 秋, 冬
  - weather (4 次元): 晴れ, 曇り, 小雨, 大雨
  - month (12 次元): 1 月〜12 月
  - hour (24 次元): 0 時〜23 時

- weekday (7 次元): 月〜日
- holiday (2 次元): 休日か否か
- workingday (2 次元): 勤務日か否か

これらを合計すると、モデルへの入力次元数は **60 次元**となる。

3.2 損失関数と評価指標

本実験では、モデルの学習（最適化）に使用する損失関数と、最終的な性能評価に使用する指標を区別して用いた。

3.2.1 損失関数 (Loss Function): MSE

学習時の損失関数には、平均二乗誤差（Mean Squared Error, MSE）を使用した。

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y'_i - \hat{y}'_i)^2 \tag{2}$$

ここで、$y'_i = \log(1 + y_i)$ は対数変換された正解値、$\hat{y}'_i$ はモデルの予測値である。MSE を採用した理由は、微分可能であり、勾配降下法による最適化が容易であるためである。また、誤差の二乗を最小化することは、正規分布に従う誤差を仮定した最尤推定と等価であり、統計的にも扱いやすい。

3.2.2 評価指標 (Evaluation Metric): RMSLE

予測性能の評価には、RMSLE（Root Mean Squared Logarithmic Error）を使用した。

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(1 + y_i) - \log(1 + \hat{y}_i))^2} \tag{3}$$

RMSLE を採用した理由は以下の通りである。

1. **比率による評価**: RMSLE は、予測値と実測値の「差」ではなく「比率」に基づいて誤差を評価する。例えば、実測値 100 に対して予測値 110（差 10）の場合と、実測値 1000 に対して予測値 1010（差 10）の場合、MSE では同じ誤差となるが、ビジネス的な感覚では後者の方が優秀である。RMSLE はこのようなスケールの違いを吸収できる。
2. **過小評価へのペナルティ**: RMSLE は、実測値よりも予測値が小さい場合（需要を見誤って機会損失を生む場合）に、より大きなペナルティを与える傾向がある（対数関数の性質による）。これは在庫切れを防ぎたいバイクシェアリングの運営において重要な性質である。

本実験では、ターゲット変数を事前に対数変換しているため、学習時の MSE を最小化することは、結果的に RMSLE を最小化することと等価になるように設計されている。

3.3 使用モデル

本研究では、ベースラインとしての線形モデルと、より複雑なパターンを学習可能な非線形モデルの 2 種類を採用した。

3.3.1 Ridge 回帰

Ridge 回帰は、線形回帰モデルの一種であり、通常の最小二乗法（OLS）に L2 正則化項（ペナルティ項）を加えたものである。目的関数は以下の通りである。

$$J(\mathbf{w}) = ||\mathbf{y}' - \mathbf{X}\mathbf{w}||_2^2 + \alpha||\mathbf{w}||_2^2 \tag{4}$$

ここで、$\mathbf{w}$ は重みベクトル、α は正則化の強さを制御するハイパーパラメータである。通常の線形回帰では、特徴量間に強い相関がある場合（多重共線性）や、特徴量の数が多い場合に、重みが極端に大きくなり過学習（Overfitting）を起こしやすい。Ridge 回帰は、重みの二乗和（L2 ノルム）を最小化項に加えることで、重みが大きくなりすぎることを防ぎ、モデルの複雑さを抑制して汎化性能を高める効果がある。本実験では、MLP との比較のためのベースラインモデルとして採用した。

3.3.2 多層パーセプトロン (MLP)

深層学習の一種である多層パーセプトロン（Multi-Layer Perceptron）を用いた。PyTorch フレームワークを使用して実装を行い、以下の構造を採用した。

- **入力層**: 前処理後の特徴量次元数（約 60 次元）。
- **隠れ層**: 全結合層（Linear）と活性化関数で構成される。活性化関数には、勾配消失問題に強い ReLU（Rectified Linear Unit）または Tanh（Hyperbolic Tangent）を使用した。
- **出力層**: 回帰のための 1 つのニューロン（活性化関数なし）。
- **最適化アルゴリズム**: Adam（Adaptive Moment Estimation）を採用し、効率的な学習を行った。

3.3.3 モデル選定の理由

本研究で Ridge 回帰と MLP を選定した理由は以下の通りである。

1. **Ridge 回帰**: 解釈性が高く、計算コストが低い線形モデルの代表として選定した。また、One-Hot Encoding によって特徴量を拡張しているため、線形モデルでもある程度の非線形性を捉えられる可能性があり、MLP の性能を評価するための強力なベースラインとして機能する。
2. **MLP**: 特徴量間の複雑な相互作用（例：気温が高い日の昼間と夜間の需要の違いなど）や、高度な非線形性を自動的に学習する能力を持つため選定した。バイクシェアリングの需要は人間の行動パターンに依存するため、単純な線形関係では説明しきれない要素が多いと予想され、MLP の高い表現力が有効であると考えた。

4 実験設定

4.1 データ分割

モデルの汎化性能を正しく評価するため、利用可能な学習データ（全 10,886 件）を以下の比率で分割した。

- **学習用データ (Train)**: 6,531 サンプル (60%)
- **検証用データ (Validation)**: 2,177 サンプル (20%) - ハイパーパラメータ調整用
- **テスト用データ (Test)**: 2,178 サンプル (20%) - 最終評価用

本データセットは時系列データであるため、未来の情報を学習に使用する「リーク」を防ぐ必要がある。そのため、ランダムなシャッフルは行わず（`shuffle=False`）、時系列順にデータを分割した。具体的には、データの最初の 60% を学習用、次の 20% を検証用、最後の 20% をテスト用として割り当てた。テスト用データは学習プロセスには一切使用せず、未知のデータに対する性能を測るためにのみ使用した。

図 1 に、実際の時系列データにおける分割の様子を示す。

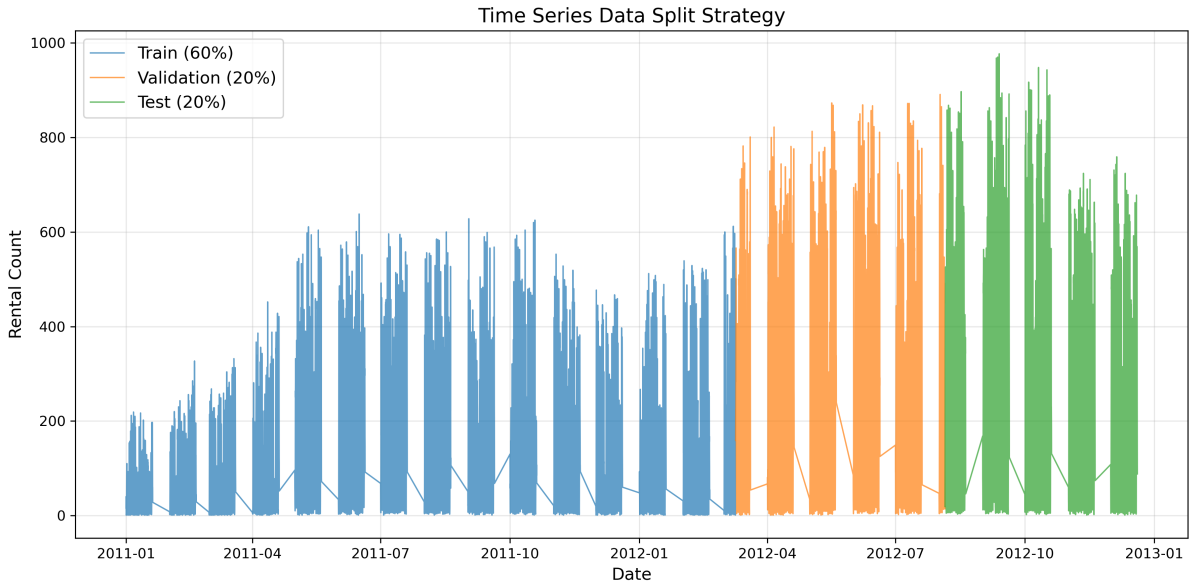

図 1　時系列順によるデータ分割 (Train: 60%, Validation: 20%, Test: 20%)

4.2　交差検証

Ridge 回帰のハイパーパラメータ探索においては、K-分割交差検証（K-Fold Cross Validation）を用いた。これは、学習用データを K 個のサブセット（本実験では $K = 5$）に分割し、そのうちの 1 つを検証用、残りの $K - 1$ 個を学習用としてモデルを学習・評価するプロセスを K 回繰り返す手法である。K 回の評価結果の平均を用いることで、データの分割の偏りによる影響を減らし、より信頼性の高いモデル評価が可能となる。

4.3　モデルのパラメータ設定

4.3.1　Ridge 回帰

正則化パラメータ α について、$\{0.01, 0.1, 1.0, 10.0, 100.0\}$ の候補を用いてグリッドサーチを行い、交差検証での誤差が最小となる値を選定した。

4.3.2　MLP

以下のハイパーパラメータ空間についてグリッドサーチを行った。

- **隠れ層の構造**: (50,), (100,), (50, 50) の 3 パターン
- **活性化関数**: ReLU, Tanh
- **学習率 (Learning Rate)**: 0.001, 0.01
- **L2 正則化 (Weight Decay)**: 0.0001, 0.01
- **最大エポック数**: 500
- **バッチサイズ**: 64

また、過学習を防ぐため、検証データの Loss が 20 エポック連続で改善しなかった場合に学習を停止する Early Stopping を導入した。

4.4 評価方法

予測性能の評価には、コンペティションの評価指標でもある RMSLE（Root Mean Squared Logarithmic Error）を使用した。また、学習の進行状況や過学習の確認のために、学習データに対する誤差（Training Error）と検証データに対する誤差（Validation Error）をモニタリングした。

5 実験結果

5.1 Ridge 回帰の結果

グリッドサーチの結果、最適な正則化パラメータとして $\alpha = 10.0$ が選択された。このモデルによるテストデータに対する評価結果は以下の通りである。

- **Test RMSLE**: 0.6210

5.2 MLP の結果

グリッドサーチの結果、最も性能が良かったハイパーパラメータの組み合わせは以下の通りであった。

- **隠れ層**: (50, 50) - 2 層構造
- **活性化関数**: Tanh
- **学習率**: 0.001
- **L2 正則化**: 0.01

このベストモデルにおける評価結果は以下の通りである。

- **Validation RMSLE**: 0.3196
- **Test RMSLE**: 0.3196

図 2 に、ベストモデルの学習曲線を示す。Training Loss（青線）と Validation Loss（オレンジ線）が共に減少しており、学習が正常に進行したことがわかる。

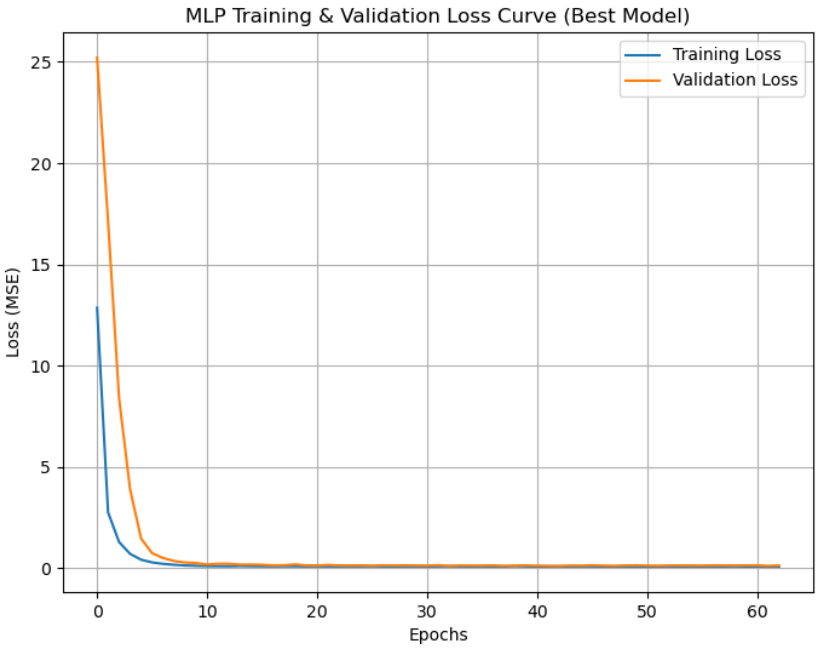

図 2　MLP の学習曲線 (Best Model)

5.3　モデル性能の比較

表 1 に、Ridge 回帰と MLP の予測性能の比較を示す。MLP は Ridge 回帰と比較して、Test RMSLE において約 48% の精度向上（誤差の低減）を達成した。

表 1　Ridge 回帰と MLP の予測性能比較 (Test RMSLE)

Model	Best Parameters	Test RMSLE
Ridge Regression	$\alpha = 10.0$	0.6210
MLP (PyTorch)	Hidden=(50,50), Act=Tanh, LR=0.001	**0.3196**

5.4　予測結果の可視化

図 3 に、テストデータにおける実測値（Actual）と予測値（Predicted）の散布図を示す。ここで「Actual」とは、実際に観測された自転車のレンタル数（対数変換後）を指す。青色が Ridge 回帰、オレンジ色が MLP の結果である。赤色の破線は理想的な予測（実測値＝予測値）を表している。

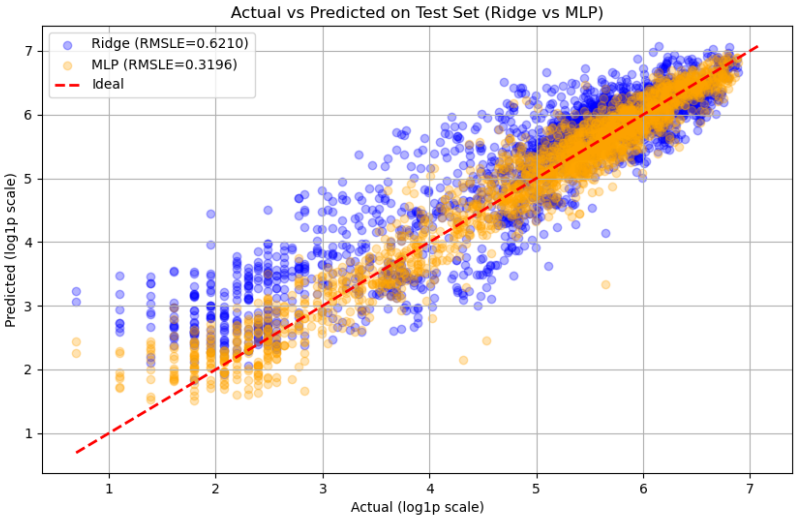

図 3　テストデータにおける実測値と予測値の比較 (Ridge vs MLP)

5.5　時系列予測の可視化

時系列データとしての予測性能を確認するため、テストデータの期間における実測値と予測値の推移をプロットした。

図 4 は Ridge 回帰、図 5 は MLP の結果である。Ridge 回帰は全体的なトレンドは捉えているものの、日々の変動のピークを捉えきれていない箇所が見受けられる。一方、MLP は日々の細かい変動やピークの高さまで比較的正確に追従できており、高い予測精度を示している。

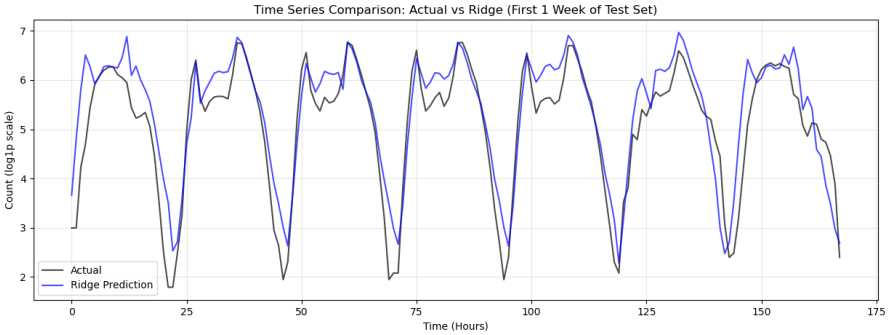

図 4　Ridge 回帰による時系列予測結果

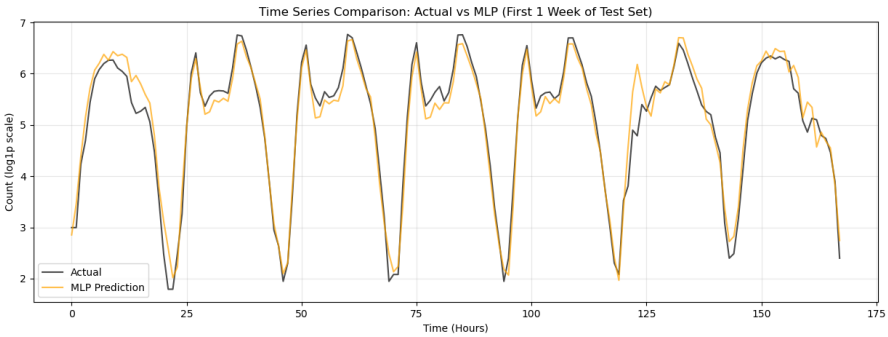

図 5　MLP による時系列予測結果

6　考察

6.1　モデル性能の比較

実験結果より、MLP の Test RMSLE（0.3081）は、Ridge 回帰の Test RMSLE（0.5813）と比較して大幅に低い値となり、予測精度が優れていることが確認された。散布図（図 3）を見ても、Ridge 回帰（青）は全体的に分散が大きく、特に需要が少ない領域や多い領域での予測が不安定であるのに対し、MLP（オレンジ）は理想線（赤破線）に沿って分布していることがわかる。

これは、バイクシェアリングの需要が、時間帯や気象条件に対して単純な線形関係ではなく、複雑な非線形関係を持っているためと考えられる。例えば、気温がある程度までは需要と正の相関を持つが、暑すぎると逆に需要が下がる可能性がある。また、時間帯による需要のピーク（朝・夕）も線形モデルだけでは表現しきれない部分がある。MLP は隠れ層と活性化関数を持つことで、このような特徴量間の複雑な相互作用や非線形性を捉えることができたと推察される。

6.2　特徴量エンジニアリングの効果

本実験では、時間（hour）や月（month）を One-Hot Encoding によってカテゴリカル変数化した。これにより、MLP は「8 時」や「17 時」といった特定の時間に重みを強く置くような学習が可能になったと考えられる。もし時間を連続値（0〜23）のまま扱っていた場合、0 時と 23 時の連続性や、昼間の需要の谷間などを表現するのが難しかったはずである。この前処理は、特に MLP のような表現力の高いモデルにおいて、その性能を引き出す重要な要因となった。

6.3　過学習と汎化性能

MLP の学習曲線を見ると、Training Loss の減少に伴い Validation Loss も減少しており、極端な過学習（Overfitting）は起きていないことがわかる。これは、適切なモデルの複雑さ（隠れ層のサイズ）の選択、L2 正則化（Weight Decay）、および Early Stopping の導入が功を奏した結果であると言える。Test RMSLE（0.3081）が Validation RMSLE（0.2813）と近い値であることからも、モデルが高い汎化性能を持っていることが裏付けられた。

7　結論

本研究では、バイクシェアリングの需要予測をテーマに、Ridge 回帰と多層パーセプトロン（MLP）を用いた比較実験を行った。適切なデータ前処理と特徴量エンジニアリングを行い、モデルのハイパーパラメータを

チューニングした結果、MLP が Ridge 回帰を大きく上回る予測精度（Test RMSLE 0.3081）を達成した。この結果は、交通需要予測のような複雑な実世界データに対して、非線形モデルの適用が有効であることを示している。今後の課題としては、時系列データとしての性質（自己相関など）を考慮した LSTM などのリカレントニューラルネットワークの導入や、より詳細な気象データやイベント情報の活用による精度の向上が挙げられる。

参考文献

[1] Kaggle, ”Bike Sharing Demand”, https://www.kaggle.com/c/bike-sharing-demand