オープンデータを用いた勾配ブースティング手法による建築工事予測モデルの検証

Using Open Data Validation of Building Construction Forecasting Model by Gradient Boosting Method

○仲川　正則[*1]，北原　英雄[*1]，加藤　万理[*1]，世古口 元伸[*1]
Masanori Nakagawa[*1], Hideo Kitahara[*1] , Mari Katou[*1], Motonobu Sekoguchi[*1]

*1 株式会社竹中工務店
TAKENAKA CORPORATION

キーワード：オープンデータ; 機械学習; ニューラルネットワーク; 都市計画
Keywords: Open Data; Machine Learning; Neural Network; Urban Planning.

1. はじめに

現在、都市が抱える課題として、人口減少や高齢者の増加、市街地の拡散等により、都市生活を支える機能の低下や地域経済の衰退などが挙げられ、社会問題となっている。国土交通省ではそのような問題に対処するために、都市機能のコンパクト・ネットワークの機能化、官民データの利活用による地域の活性化などを通じて、官民連携したスマートシティへの取り組みを行っている。その政策の一環として、各自治体や民間企業に向けて都市計画基礎調査情報[1]の利活用を推進、ガイドラインの策定・公表などを行っている。この都市計画基礎調査とは、都市計画法に基づき、都市現況及び将来の見通しを定期的に把握するための調査であり、本調査で収集される情報は Table 1 に示すような土地や建物の利用状況、例えば建物用途やその階数、構造、面積、建築年など建物に応じた詳細なデータを街区単位で集計されるものである。調査情報については現在、各自治体、都市あるいは民間企業、個人など幅広く共有できる状態にすることを念頭にフォーマットが定義されている。これらオープンデータを利活用し具体的なユースケースを立案するための手順書などの整備もされ始めており、今後具体的な都市整備計画などへの利用が活発化することが予想される。

都市計画基礎情報を活用することで例えばある都市における最適な都市基盤整備のための建物の地域上でのレイアウトやその都市における成功事例データを別の都市で展開した場合に基盤整備をどのように行えば最適化されるのか等を予測することが期待できる。今回の検証では将来的な各地域のオープンデータ蓄積、またそれを活かした都市基盤整備計画やユースケース立案のために、都市計画基礎情報を用いた建築工事予測を行うことを目的とした。ただし現状、都市基盤整備計画情報は全国規模でオープンデータとしては公開されていないため、今回は当社の建築工事データを用いた。最終的に、建築工事に関するあらゆるオープンデータを活用し、官民様々な都市機能に対するユースケースに対する予測モデルを構築し、幅広い施策に適用することで、都市機能最適化への貢献を目的としている。また、建築工事トレンド予測を行う上では建物情報だけではなく、その地域における経済状況も加味しなければならず、各地域の景況動向データを付随することでその補完を行った。また、すべてのデータセットについて各々のデータ項目が収集できていることが望ましいが、一部の地域や期間でデータが欠如しているケースなども考えられる。そのため、予測に対する手法については、データセット内に欠損値が存在する場合かつ欠損そのものに意味がある場合にも対応可能な勾配ブースティング決定木(以下 GBDT)の各手法を用いた。GBDT は Kaggle などの予測モデリングコンペティションなどで上位入賞モデルの過半を占め、どのようなドメインの分析対象にも比較的精度が安定しており、社会実装を目的とした際に有用である。また、予測モデルに対する学習方法については、通常では教師あり学習を用いることが多いが、予測対象が欠如したデータに対する教師なし学習による事前学習を併せた半教師あり学習も併せて検証を行った。この事前学習を用いた予測モデルを活用しその精度を上げることで、ある期間にデータ欠損があるような状況下や、ある都市では存在しているデータ項目が別の都市ではない場合でも対応可能にする。これは将来的な様々なオープンデータの形式に対応した予測モデルを目指すためである。

Table 1. 都市計画基礎調査情報

現況区分	項目
土地利用現況	用途別土地利用面積
建物利用現況	建物用途、階数、構造、 建築面積、延床面積、 建築年、耐火構造種別

2. データについて

2.1. 使用データ

本検証で使用するデータは Table 2 に示す統計データおよび建築工事データを用いた。各データとも地域別である。

1.景況動向データ

2.建築工事データ

1 については内閣府が公開している景気動向指数(DI: Diffusion Index)および地域別支出総合指数(RDEI: Regional Domestic Expenditure Index)のうち、地域別消費総合指数を用いた。2.については建築工事の建物単位データを用いた。

2.2. データの内訳

景気動向指数は景気の波及が住宅関連、金融関連などの部門にどれだけ波及したかを示す値である。0~100 の間で変動し、50 を上回る場合は景気の拡張期、50 を下回る場合は後退期の目安となる。Table 3 に示すような項目となり、全国、その他地域ブロック毎に算出されている。Figure 1 に 2017 年から 2020 年 8 月までの DI の値を示す。2017,2018 年は概ね景気拡大期と衰退期の上下で推移しているが、2019 年以降衰退期に入り、2020 年では COVID-19 による経済への影響から大幅に減少していることがわかる。今回の検証では 2016~2019 年の期間で地域ブロック別月次現状値データを年度平均した値を用いた。

地域別支出総合指数については、地域別消費総合指数を用い、これは財、住宅、サービスからなる 44 の指標ベースでそれぞれの基準支出額を算出し、季節調整をかけた後に算出、各基準の月次ベースの支出額を 2012 年 1 月の値を 100 とし導出したものである。Table 4 に示すような項目となり、こちらも全国およびその他地域ブロック別に算出されている。Figure 2 に示すように、2019 年までは、月次変動に周期性が見られ年次単位で同様の推移を示している。2019 年については、ピーク値も減少しており、当年後半にかけて前年を下回るような傾向が見られている。また 2020 年以降は COVID-19 の影響を受けていることは景気動向指数と同様の傾向である。こちらも 2016~2019 年の期間で地域ブロック別月次現状値データを年度平均した値を用いた。これら指数データと 2015 年から 2019 年までの建築工事データの建物単位毎のデータと結合した。建築工事データの内訳は、都市計画基礎調査情報の項目にある土地利用面積、建物用途、階数、構造、建築面積、建築年などになる。また指数データを含め標準化を行った。本来であれば GBDT の各手法について標準化や正規化などの処理は不要であるが、後に説明する教師なし学習におけるニューラルネットワークによる事前学習時に、データ項目ごとの更新の偏りを防ぐためである。目的変数は次期建築工事発生有無とした。検証データについては、k-分割交差検証(k=5)により訓練データと検証データの分離とした。

Table 2. 本検証で用いたデータ

統計データ	
景況動向データ	地域別景気動向指数(DI)
	地域別支出総合指数(RDEI)内の地域別消費総合指数
建物別データ	
建築工事データ	建築工事の建物単位データ

Table 3. 地域別景気動向指数例

	北海道	関東	東海	近畿	九州
2019/1	49.3	46.0	44.6	45.8	44.5
2019/2	51.0	46.3	45.6	48.0	46.8

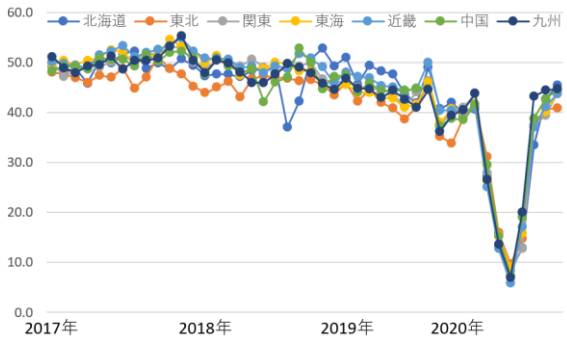

Figure 1. 地域別景気動向指数の推移(2017～2020 年度)

Table 4. 地域別消費総合指数例

	北海道	関東	東海	近畿	九州
2019/1	104.55	101.37	101.09	108.44	101.59
2019/2	107.65	102.42	100.75	107.40	102.17

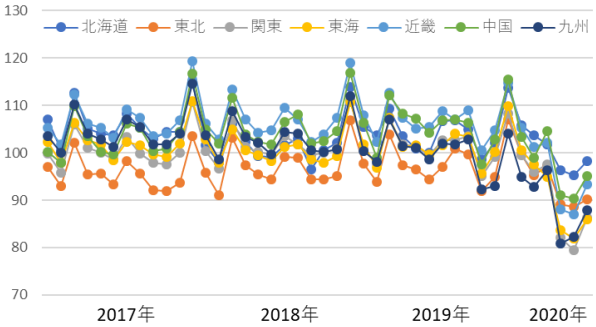

Figure 2 地域別消費総合指数の推移(2017～2020 年度)

3. 分析手法

3.1. GBDT

今回は複数の弱学習器を加法的に追加する学習方法である GBDT を用いた。種々の統計、機械学習モデルの中でも社会実装の観点で決定木における要因分類の可視化など結果に対する説明可能が容易である。今回の検証におい

て建築工事の増減に対しどのような要因が影響を与えているかを明らかにするため GBDT において種々のアルゴリズムがあるが、XGBoost[2], LightGBM[3],CatBoost[4]を用い比較実験を行った。この 3 つのアルゴリズムはその他アルゴリズムと比べ研究・実績結果などナレッジも充実しており、今回のようなテーブルデータに強くかつ、計算速度が速いなどのメリットがあるためである。

3.2. 半教師あり学習

本検証では、教師あり学習の他に半教師あり学習としてラベル無データを用いた Auto Encoder による教師なし学習で得られた隠れ層(中間層)による事前学習による訓練データを用いることで教師あり学習と比べ予測精度の優位性があるかの検証を行った。

3.3. 分析手法

今回の分析手法のシーケンスは Figure 3 の通りとなる。

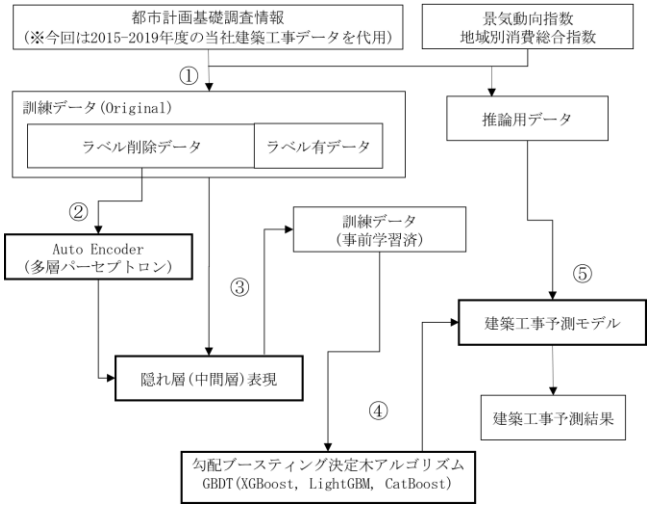

Figure 3. 本分析手法シーケンス

以下に本分析手法のステップを記載する。

① 建築工事データと景気動向指数と地域別消費総合指数の年度別/地域別指数データを結合し訓練データを生成する。また、推論用データとして訓練データの次年度データを別に分類する。

② あらかじめラベルデータを削除した訓練データを用意し、多層パーセプトロン構成の Auto Encoder を通して教師なし学習による隠れ層表現を得る。

③ 全ての訓練データを用いて②で作成した隠れ層へのインプットを行い、新たに生成された特徴量と併せた教師なし学習による事前学習済み訓練データを得る。

④ ③で得られた訓練データを GBDT アルゴリズムへのインプットを行い学習させ、建築工事予測モデルを出力する。

⑤ 推論用データを④で得られた予測モデルに入力することで予測結果を得る。

4. 検証結果

4.1. 設定

GBDT および Auto Encoder のハイパーパラメータについては、学習時に Optuna[5]を利用し最適化を行った。次に事前学習済訓練データ作成のため隠れ層の生成を行う。そのため、元の訓練データの正解ラベルから 75%の正解データを除去したデータを用いて Auto Encoder による事前学習を行った。また LightGBM および CatBoost においては、説明変数のうち順序性のないものに関し、カテゴリ変数として定義し学習を行った。

4.2. 予測モデルの評価

本項では教師あり学習、Auto Encoder による教師なし事前学習により生成した訓練データを用いた半教師あり学習の比較検証を行う。検証データを適用した結果について、それぞれの対数損失(Log Loss)、Precision、Recall、AUC を評価指標として比較を行った。Table 5, Table 6, Figure 4, Figure 5 のように各 GBDT において損失値はほぼ変わらず、正答率は若干半教師あり学習が高く、この傾向は正解ラベルの削減割合を変えても傾向は変わらなかった。3 つのアルゴリズムについては LightGBM が他 2 つを若干上回る結果となった。XGBoost に比べて LightGBM の損失が低い理由については決定木の分岐の扱い方において XGBoost では Level-wise という深さ単位で分岐を増やしていくのに対し、LightGBM では Leaf-Wise という葉単位に分岐するする手法を用いており葉毎に目的関数を減少させる分割が可能になるため推論精度が高くなっていると考えられる。また、CatBoost においては、教師あり学習では、他の 2 つのアルゴリズムと遜色ない結果であったが、半教師あり学習の場合は対数損失の値が他と比べ比較的上昇した。隠れ層による事前学習で得られた特徴量には明示的なカテゴリ変数の定義が困難なため、本アルゴリズムの特徴であるカテゴリ変数に対する最適化の効果が説明変数全体の増大に対し弱まったためであると考えられる。ただし、教師あり学習では LightGBM と遜色なく、今後データ拡張時にカテゴリ変数の割合が増えた場合に優位性が出る可能性もあり、かつ説明変数及び目的変数内で相関関係が強い変数がある場合は精度が向上するため、併せて検証を進めていきたい。

次に LightGBM においても CatBoost と同様に明示的にカテゴリ変数として定義し学習させた結果を Table 7 に示した。教師あり、半教師あり学習ともに対数損失は上昇する結果となった。これはカテゴリ変数が過剰適合し、過学習に陥った結果であると考えられ、様々な研究にて同様の現象が報告されている。また、それらカテゴリ変数に対し label-encoding ではなく one-hot-encoding により説明変数として置き換えた結果を Table 8 に示した。教師あり学習に関しては若干の損失減少が見られたが、半教師あり学習では、隠れ層による事前学習の際に次元数が増大した影響を

解消できず最終的な対数損失は Table 6 で示したカテゴリ変数を数値データとして定義したケースを下回ることはなかった。

Table 5. 教師あり学習における
各 GBDT の対数損失、適合率、再現率

Algorithm	Log Loss	Accuracy	Precision/Recall
XGBoost	0.5904	0.6562	0.7907/0.7013
LightGBM	0.5782	0.6736	0.7606/0.7012
CatBoost	0.5834	0.6771	0.7819/0.7142

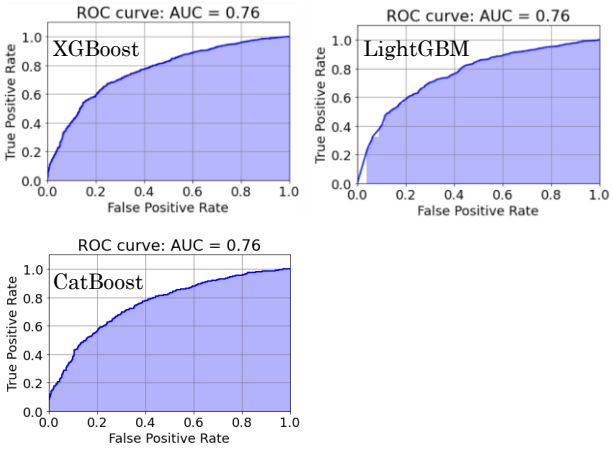

Figure 4. 教師あり学習における各 GBDT 手法の
ROC 曲線および AUC

Table 6. 半教師あり学習における
各 GBDT の対数損失、適合率、再現率

Algorithm	Log Loss	Accuracy	Precision/Recall
XGBoost	0.6024	0.6714	0.7721/0.7143
LightGBM	0.5818	0.7058	0.7799/0.7077
CatBoost	0.7544	0.6775	0.7538/0.7338

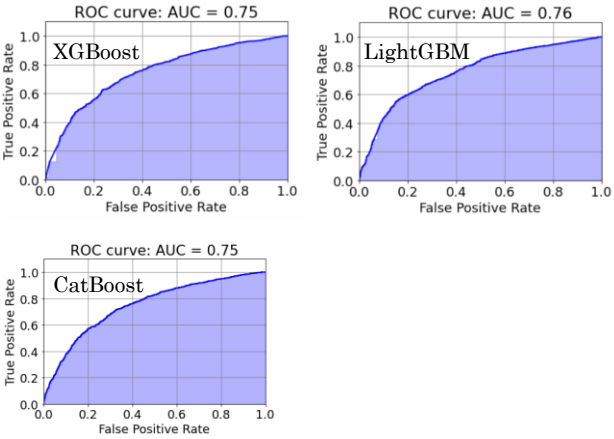

Figure 5. 半教師あり学習における各 GBDT 手法の
ROC 曲線および AUC

Table 7. LightGBM カテゴリ変数定義による
対数損失、適合率、再現率

	Log Loss	Precision	Recall
教師あり	0.5815	0.7721	0.7143
半教師あり	0.6007	0.7725	0.8759

Table 8. LightGBM カテゴリ変数の one-hot encoding による対数損失、適合率、再現率

	Log Loss	Precision	Recall
教師あり	0.5813	0.7721	0.7143
半教師あり	0.6143	0.7638	0.6494

5. まとめ

本検証では、地域毎の景況動向データを用いつつ、各種勾配ブースティングアルゴリズムや学習方法の違いにより、将来の建築工事発生予測の推論結果や精度の違いを示した。ラベルデータの大部分がないケースを想定し半教師あり学習による精度向上も試みたが、次元数増大による推論精度の低下など課題が残った。ただし、オープンデータとして全国の街区エリアすべてを網羅するかは現状不明であり、かつユースケースによっては正解データが一部しか存在しないケースについても対処する必要があることから、ラベル無データを活用した予測分析は試行する必要性がある。また、各地域の建築工事にまつわる要因は今回用いた景況動向データだけでなく、国内・国際政治情勢、経済動向などの指数や、地域別の災害や都市計画なども考慮にいれなければならない。こういった多軸要素を加えながら、オープンデータとして都市計画基礎情報の各地域の街区レベルまでのデータが入手できれば、特定の地域にスコープを当てたより細かな予測、分析活動も可能となり、先に挙げた都市機能の最適化に向けたユースケースの立案などの一助になると考えられる。

[参考文献]

1)国土交通省 HP 都市計画基礎調査情報の利用・提供ガイドライン https://www.mlit.go.jp/common/001282175.pdf

2)Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.

3) Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems. 3149–315

4) Hinton, G.E. and Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. Science, 313(5786):504–507, 2006.

5) Preferred Networks, Inc https://preferred.jp/ja/projects/optuna/