ISPOR 2023
MSR17

Imputation of Line-Level Medical Costs Using a Gradient Descent-Boosting Regressor Model

Bill Woywod, Ron Katriel — Komodo Health, New York, NY and San Francisco, CA

Introduction

Health-insurance claims records that include financial fields such as the allowed amount are critical to a variety of research disciplines, including health economics research, health services research, outcomes research, and actuarial sciences. Due to the protected nature of payer-provider rate negotiations, financial data is commonly redacted from many commercially available claims data sources. Establishing a reliable methodology for imputing financial fields holds the promise of greatly expanding the universe of claims data available to researchers.

Objective

The objective of this study was to use machine learning to accurately impute missing service-level medical claim allowed amounts.

Methods

Line-level allowed amounts were predicted using a gradient descent-boosting regressor machine learning (ML) model. This decision tree ensemble algorithm first developed by Friedman in 1999 combines the potential for high predictive accuracy with a high degree of flexibility (e.g., it can optimize different cost functions and be extensively tuned). The training dataset included 26 million claim-line records from Komodo's Healthcare Map™ representing insurance segments in which reimbursement is subject to rate negotiation: commercial, Medicare Advantage, and managed Medicaid.

The model was built on a set of 24 independent variables designed to capture factors such as bargaining power that are capable of driving variation in negotiated rates. The variables were used across the following domains:

- Payer characteristics and market power
- Provider characteristics, market power, and practice costs
- Local market characteristics and demographics
- Category of service

Rather than predicting the cost of a service directly as a dollar amount, the model takes a novel approach to structuring the dependent variable that predicts whether a claim line is reimbursed above or below the national median amount for that particular service. This dependent variable, called the "payment ratio," takes on a value of 1 for a claim that is reimbursed at the median value for that type of service, 1.1 for a claim priced 10% above the median, and so on. This approach has two principal benefits:

1. It improves the model's generalizability, making it capable of pricing claims irrespective of the procedure code, and
2. It improves the model's performance. The model was validated on an unseen sample of 2.7 million records of withheld data to guard against model overfitting and target leakage.

The model was then optimized across several hyperparameters, including the number of estimators, maximum tree depth, learning rate, and reg alpha (L1 regularization term on weights).

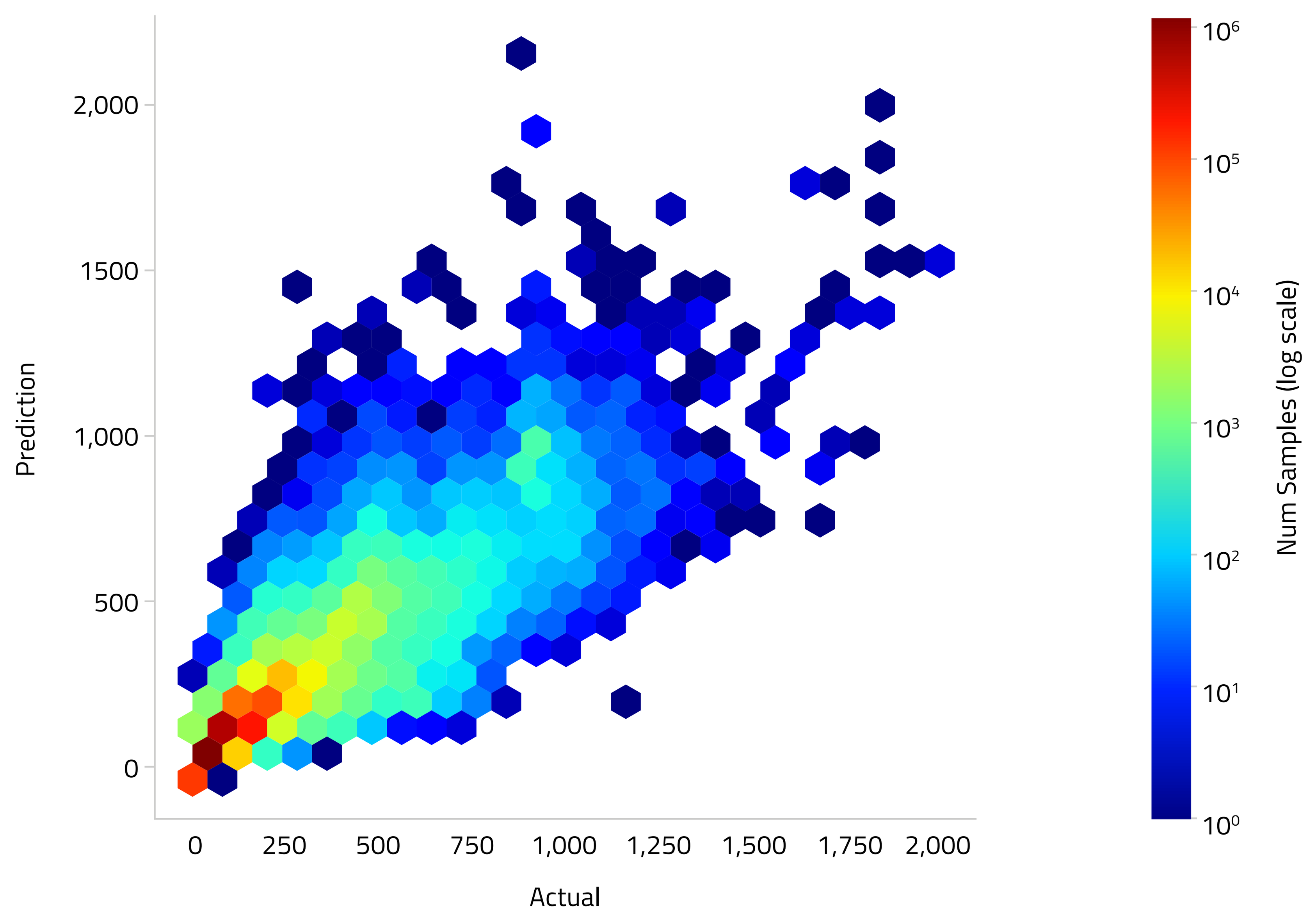
Results

The model exhibited strong performance upon validation, with an R2 value of .87 and a mean absolute error (MAE) of $13.83. Due to the generalizable, code-agnostic design, the model was able to estimate allowed amounts for 99.5% of the records in the validation sample.

Table I: Model Performance Measures

Model	Statistic
R^2 Score	0.87
MAE	$13.83
RMSE	$31.06

Chart 1: Hexbin of Actual vs. Predicted Allowed Amounts



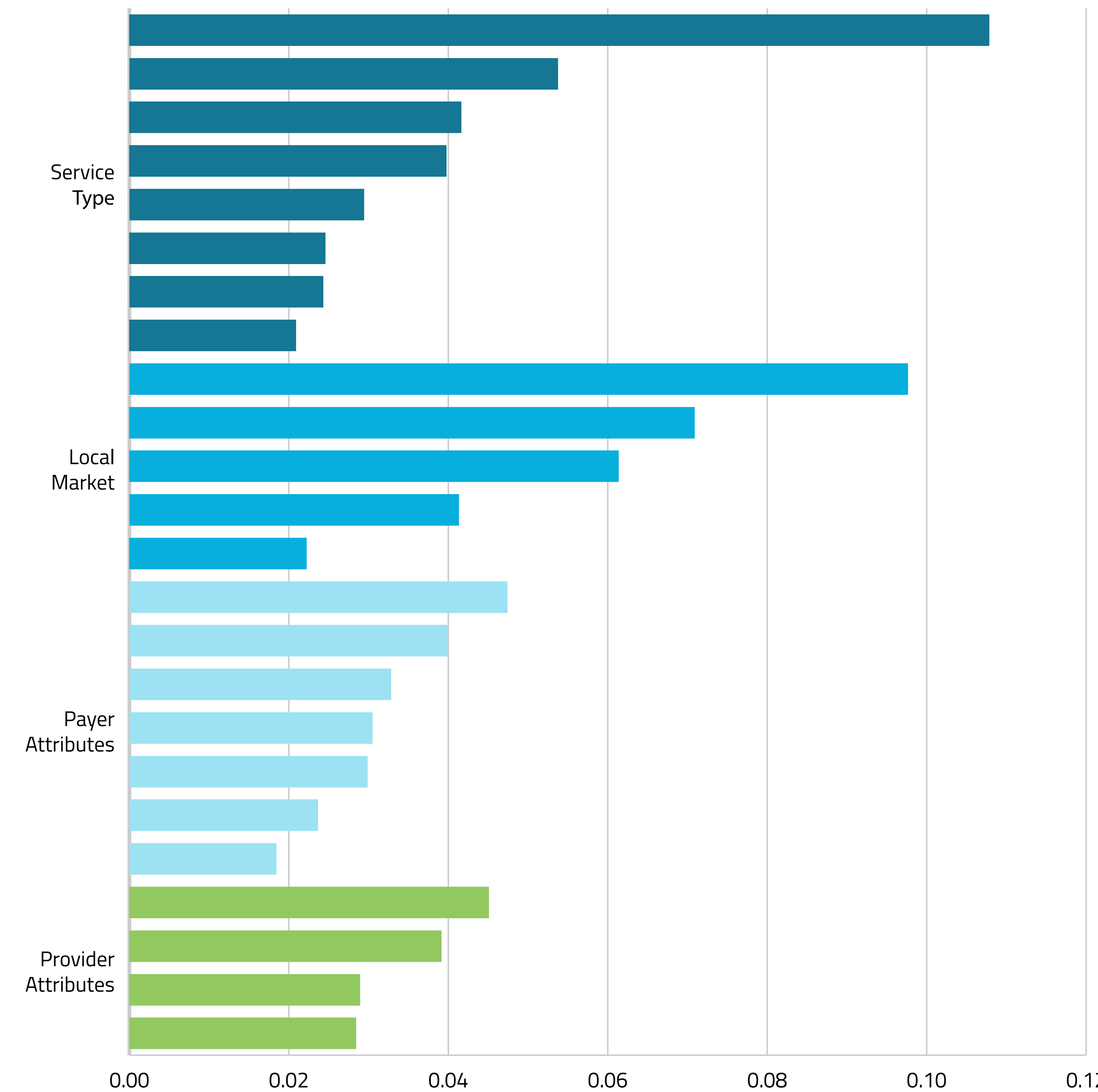
When calculated across the payer channel, performance measures are highest for the Medicare Advantage segment. Rates negotiated by Medicare Advantage plans may be anchored against Medicare fee-for-service (FFS) rates, which are well understood, publicly available, and have lower variance than commercial negotiated rates, which may offer an explanation for the high performance and predictability observed in that segment.

Table 2: Model Performance Measures by Payer Channel

Payer Channel	R^2	MAE	RMSE
Commercial	0.837	$16.84	$41.34
Medicare Advantage	0.940	$7.30	$23.30
Managed Medicaid	0.836	$14.38	$30.78

When grouped into like categories, independent variables related to the type of service had the largest impact on model predictions, followed by local market, payer, and provider characteristics.

Chart 2: Model Feature Performance, Assessed Using Mean Decrease in Impurity



Conclusion

The gradient descent boosting regressor, coupled with a robust and carefully designed independent and dependent variable table structure, is capable of producing accurate estimates of service-line level medical allowed amounts across a large number of procedure codes and three payer types.

References

[1] Friedman, JH. Stochastic Gradient Boosting. 1999. Retrieved October 27, 2022. https://jerryfriedman.su.domains/ftp/stobst.pdf

Scan here to download poster or inquire for more info.

komodo™