SML 301 Semester Project Report

Yoselin Chavez

Princeton University

SML301

Professor Derek Sollberger

04/05/2025

Author Note

Honor Pledge: I pledge my honor that I have not violated the honor code during this examination.

Maria Yoselin Chavez

Introduction

The rising costs and inconsistent pricing in the U.S. healthcare system have resulted in a growing demand for transparency in hospital billing practices. With recent federal mandates requiring hospitals to publish their pricing data, the availability of such datasets has opened up opportunities for data driven analysis. This project focuses on building predictive models to forecast hospital prices using machine learning techniques, drawing upon publicly available data provided by the Centers for Medicare & Medicaid Services (CMS). It will be shown how to develop accurate predictive models for healthcare costs, as well as demonstrate how algorithmic tools can enhance pricing transparency.

Our primary objective here is to explore and evaluate the effectiveness of various machine learning algorithms in modeling hospital price data. We will implement and compare four regression models to predict log transformed procedure prices, such as Random Forest Regressor, K-Nearest Neighbors, Linear Regression, and PyTorch Neural Network. These models are trained on features such as hospital location, procedure type, and diagnoses related group codes, enabling us to assess the predictive accuracy and interpretability of different approaches. The data comes from the DoltHub Hospital Price Transparency repository, merging the three tables of prices, hospitals, and cpt_hcpps to get procedure costs by hospital, facility metadata, and procedure code descriptions.

This work's purpose is to serve both technical and practical purposes. From a technical perspective, there's the opportunity to apply and compare supervised learning methods in real world settings. From a practical standpoint, there's the opportunity to contribute to the ongoing efforts to enhance transparency and accountability in healthcare pricing. Ultimately, this project

situates itself at the intersection of data science and healthcare policy, seeking to provide insights that could inform both model development and public discourse.

Data Description

The data used consists of a hospital pricing dataset loaded from DoltHub's Hospital Price Transparency repository, using Python's doltcli library. This repository contains federally mandated hospital pricing disclosures structured across tables of prices, hospitals, and cpt_hcpcs. These are procedure level pricing data, facility metadata, and clinical procedure terminology code descriptions, merged using the hospital identifier and procedure identifier, with a random sampling of 10,000 records for retained analysis.

Code Excerpt:

```python
# Load data from Dolt
def load_data_from_dolt():
    """Load data from Dolt repository, merge relevant tables"""
    try:
        repo = dolt.Dolt("hospital-price-transparency")
    except:
        repo = dolt.Dolt.clone("dolthub/hospital-price-transparency", "hospital-price-transparency")

    # Load hospitals data
    repo.sql("SELECT * FROM hospitals", result_file = "hospitals.csv")
    hospitals = pd.read_csv("hospitals.csv", dtype = {'npi_number': str, 'zip_code': str})

    # Load prices data
    repo.sql("SELECT * FROM prices", result_file = "prices.csv")
    prices = pd.read_csv("prices.csv", dtype = {'npi_number': str, 'code': str})

    # Load CPT/HCPCS codes
    repo.sql("SELECT * FROM cpt_hcpcs", result_file = "cpt_hcpcs.csv")
    procedures = pd.read_csv("cpt_hcpcs.csv", dtype = {'code': str}, low_memory = False)

    # Merge datasets
    df = pd.merge(prices, hospitals, on = 'npi_number', how = 'left')
    df = pd.merge(df, procedures[['code', 'short_description']], on = 'code', how = 'left')

    df = df.sample(10000, random_state = 42)
    return df
```

The dataset contains 180,690 rows and 18 columns. Each row corresponds to a specific hospital service, and it's provided by a given hospital payer type. The features are a mix of categorical and numerical variables, representing information about the hospital system, service type, prices, payer classification, and market characteristics.

Key Variables Include:

- facility_name: Name of individual hospital.
- system_name: Health system to which the hospital belongs.
- service: Type of medical procedure or service rendered.
- payer: Classification of payer (e.g., Medicare, Medicaid, Commercial).
- price: Reported negotiated price for the service.
- price_percent_of_medicare: A normalized measure of price relative to Medicare benchmarks.
- market_name, market_state, and market_hrr: Geographical and referral region identifiers.
- hospital_type: Indicates whether the facility is acute care, critical access, or another type.
- ownership_type: Non-profit, for-profit, or government-owned classification.
- discharges and net_patient_revenue: Proxy indicators of hospital scale and financial capacity.

The dataset shows considerable variability in service prices. The 25th percentile of commercial prices is approximately $597, while the 75th percentile reaches $2,117. The high variance reinforces the value of machine learning methods that can handle nonlinear relationships and heterogeneity across cases. Many categorical features have high cardinality, such as system_name and service, which require appropriate encoding for machine learning applications. Variables such as price_percent_of_medicare provide a standardized reference point that supports comparative pricing analysis.

For the data cleaning and preprocessing, there were several steps involved. The first involved handling the missing data, meaning observations with missing values in essential fields such as price or payer type were removed using listwise deletion. One-hot encoding was also applied to key categorical variables such as payer, ownership_type, and hospital_type. And due

to the right-skewed distribution of prices, the decision was made to apply log transformation to the price variable to stabilize variance and enhance model performance. FInally, redundant or low-variance variables were excluded from the modeling pipeline after an initial exploratory analysis.

The dataset's structure makes it well suited for supervised learning, particularly regression models. With the array of predictors and significant variation in target values (prices), machine learning models can leverage both statistical patterns and structural features to forecast outcomes. However, the presence of high cardinality categorical variables and outliers require careful feature engineering and validation, which can be addressed in the modeling section.

Literature Review

Healthcare cost prediction has become a research domain of great significance in the United States as concerns around cost transparency, affordability and efficiency rise. Growing availability of public data has allowed data-driven solutions to emerge, with approaches focusing on aspects such as irregularities, forecasting costs, and aiding policy decision making. This literature review seeks to survey several recent works that utilize machine learning for healthcare cost prediction, with particular focus on hospital charge forecasting. Covered will be contributions related to deep learning, regression models, and the policy implications.

Deep Learning

Looking at deep learning in health care cost prediction, it becomes apparent the potential of deep learning applied to electronic health records, for instance, when it comes to predictive tasks in healthcare. This study demonstrates how deep neural networks (DNNs) outperform traditional models in predicting clinical events, discharge diagnoses, and patient mortality using large scale datasets. This is particularly relevant to this project's focus, as it uses a neural

network model to predict hospital pricing data based on structured attributes such as diagnoses related group (DRG), procedure codes, and provider information. Scalability and accuracy of the deep learning framework presented in this paper shows how such architectures can be applied to hospital price prediction tasks, even though our particular dataset does not consist of the same granular EHR level data.

It's also notable that Rajkomar et al. notes the importance of feature embedding and temporal modeling, which could be adapted to enhance predictive performance of hospital billing systems. While our approach is simpler in terms of architecture and feature complexity, this study justifies our use of neural networks in modeling high-dimensional and nonlinear healthcare data.

Traditional Models: Regression and Decision Trees

Deep learning models are showing increasing promise in healthcare contexts, but traditional ML models still remain competitive and oftentimes continue to provide greater interpretability. In the study conducted by Bertsimas et al. (2008), the authors develop algorithmic models for healthcare cost prediction using regression, decision trees, and clustering techniques. It's demonstrated that these methods can yield high accuracy predictions and find cost driving patterns in the data. Their work supports our implementation of models such as linear regression, decision trees, and K-nearest neighbors (KNN), as well as neural networks. Each of these models have their own distinct strengths, such as linear regression providing interpretability, decision trees highlighting feature importance, and KNN capturing local structure in the data.

The emphasis on model comparison in the work presented by Bertsimas et al. mirrors our own methodology, where the aim is to assess the performances of different algorithms based on

RMSE, R-squared, and other evaluation metrics. This comparative approach is essential to selecting the most suitable model given the data's characteristics, such as multicollinearity, heteroscedasticity, and sparsity.

Systematic Reviews and Modeling Best Practices

Panagiotou et al. (2022) conducts a systematic review of multivariable prediction models for healthcare spending, evaluating a wide range of ML and statistical approaches. The findings of this study indicate that while advanced methods such as gradient boosting and DNNs can improve performance, simpler models can often perform comparably. This insight serves to reinforce the rationale for including both linear and nonlinear models in our analysis.

The review also identifies gaps in feature selection, validation techniques, and external generalizability across many published studies. Our work aims to address these concerns by implementing cross-validation, careful hyperparameter tuning, and using standardized datasets (CMS) to ensure replicability. Furthermore ,Panagiotou et al. highlights the need for transparent reporting of model assumptions and limitations, which have been incorporated into our discussion and conclusion sections.

Hospital Pricing and Transparency

Beyond methodological insights, it's crucial to understand the policy context underpinning hospital pricing. Bai and Anderson (2015) analyze the markup ratios across U.S. hospitals, finding extreme variability in pricing that can't be justified through patient complexity or cost inputs. Their findings underscore the opaqueness of hospital billing systems, as well as the potential for data-driven models to uncover and potentially mitigate unjustified price variation.

Incorporating Bai and Anderson's work frames the project as part of a broader dialogue on healthcare justice and market regulation. By predicting hospital prices with a degree of accuracy, we can contribute to the transparency goals promoted by federal mandates that require hospitals to publish chargemaster data. Although predictive modeling on its own can't rectify structural inequities, it can equip stakeholders with actionable insights.

Applications to Hospital Operations

While the primary focus of our research is on pricing prediction, insights from studies in adjacent areas can also prove useful. Huang et al. (2021) conducts a scoping review of ML applications in hospital readmissions predictions, shedding light on data preprocessing challenges, variables importance assessments, and outcome calibration strategies relevant to our task. The parallels in our data structure, such as hospital identifies, clinical codes, and discharge outcomes, underscore the transferability of these techniques.

Further, Huang et al. argues that ML tools should serve not only to predict outcomes, but also to inform strategic decisions in hospital operations. Our view is similar, in that price prediction models can ultimately inform audits, competitive benchmarking, and value-based care initiatives. By synthesizing predictive performance with interpretability and policy relevance, our project seeks to move beyond black-box predictions toward more responsible ML integration in healthcare settings.

Ultimately, the literature supports a diverse modeling toolkit for healthcare cost prediction, from interpretable linear regressions to expressive neural networks. Prior research has validated the use of deep learning for EHR based prediction, traditional models for cost regression, and systematic comparisons for methodological rigor. Studies in healthcare pricing

have also highlighted the urgent need for transparency and accountability, offering a compelling justification for predictive approaches.

Our project draws upon and extends this research tradition by applying multiple ML models to predict hospital pricing using public CMS datasets. In doing so, we bridge technical development and social impact, contributing not only to the field of supervised learning, but also to the broader healthcare transparency movement.

Methodology

Feature Engineering

Building on the cleaned dataset, new features were created to enhance predictive performance. From procedure codes, code_prefix (alphabetic prefix) and code_numeric (numeric portion) were extracted from the procedure codes, intended to capture groupings of procedures and their hierarchical structure. Price level statistics were also computed, grouped by procedure code to provide contextual information about typical price levels and variation.

Model Training

Three traditional regression models were trained, Random Forest Regression, K_nearest Neighbors, and Linear Regression. For each model, a pipeline was constructed that incorporated preprocessing using ColumnTransformer and Pipeline utilities from scikit-learn. The dataset was also randomly split into 80% training and 20% testing subsets using train_test_split. PyTorch based feedforward neural network was also implemented, using hidden layers with ReLU activations.

Each model was evaluated using standard metrics, meaning Mean Squared Error, Mean Absolute Error, and R^2 score. Performance differences were analyzed to understand model

strengths and weaknesses, especially when it comes to handling nonlinearity and high cardinality categorical features.

Hyperparamter Usage

To improve the model performance and generalizability, hyperparameter tuning was performed for Random Forest and neural network models, selecting reasonable defaults for K-Nearest Neighbors (KNN). Hyperparamters were selected based on cross validation and practical performance heuristics.

Random Forest

GridSearchCV was used for Random Forest Regressor with a 3 fold cross validation to optimize three key hyperparameters, n_estimators, max_depth, and min_samples_split. Variable n_estimators represent the number of trees, mx_depth represents the maximum depth of each tree, and min_samples_split represents the minimum number of samples required to split an internal node. It's been suggested in previous literature that limiting tree depth helps control model complexity, and min_samples_split > 2 encourages generalization in deeper trees.

K-Nearest Neighbors (KNN)

For KNN, n_neighbors = 5 was used as a standard starting point. The algorithm's poor performance in early evaluations, however, resulted in the choice not to pursue additional tuning, so further optimization could be explored in future work.

Neural Network (PyTorch)

A fully connected feedforward neural network in PyTorch was implemented as well. This included Input Layer, Hidden Layer 1 with 64 neurons and ReLU activation, Hidden Layer 2 with 32 neurons and ReLUE activation, and the Output Layer with 1 neuron. Training was done with Optimizer Adam consisting of learning_rate = .001, loss function of Mean Squared Error,

Epochs of 50, and the entire dataset as a batch size with batch gradient descent. These values were selected based on practical guidelines for small to medium sized tabular data, and learning rate of .0001 was chosen for its stability with Adam. The network width of 64 to 32 was also designed to capture nonlinear interactions without overfitting, however, there was no extensive tuning of layer size, dropout regularization, or learning rate schedules, which leaves room for future research.

Performance metrics suggest that the neural network was trained successfully, though generalization was rather poor, given the $R^2 = -4.02$. This could be due to overfitting, or architectural mismatch with the feature structure. Future work could incorporate techniques, such as validation based early stopping, dropout, and grid/random search for hyper parameters such as learning rate, batch size, and layer depth.

Code Excerpt:

```
# Train PyTorch Neural Network
class NeuralNetwork(nn.Module):
    def __init__(self, input_size):
        super(NeuralNetwork, self).__init__()
        self.fc1 = nn.Linear(input_size, 64)
        self.relu1 = nn.ReLU()
        self.fc2 = nn.Linear(64, 32)
        self.relu2 = nn.ReLU()
        self.fc3 = nn.Linear(32, 1)

    def forward(self, x):
        x = self.fc1(x)
        x = self.relu1(x)
        x = self.fc2(x)
        x = self.relu2(x)
        x = self.fc3(x)
        return x
```

Results

The primary objective was to predict the log transformed price variable using the hospital level, service level, and market specific features available. Evaluation metrics included MSE, MAE, and R^2, which allowed for assessment on both precision and explanatory power of each predictive model. The Random Forest Regressor demonstrated the strongest performance, with MSE = 0.2872, MAE of 0.2010, and $R^2 = 0.9258$. The tuned Random Forest Performance

showed even better results, with a Best MSE of 0.2515, which is improved from the untuned value of MSE = 0.2872. This suggests the model is highly effective in capturing nonlinear relationship and interactions amongst variables such as procedure code, payer classification, and hospital system. The use of tree based ensembles likely contributed to the robustness of the model in presence of high-cardinality categorical variables and price variability across service types.

Alternatively, KNN had much lower performance with MSE = 1.4903, MAE = 0.8488, and R^2 = 0.6149, meaning that while the model could explain some variation in log prices, it was far less accurate than Random Forest. It's likely that one limited of the model is the KNN's sensitivity to the curse of dimensionality, particularly when it comes to handling one-hot encoded categorical features.

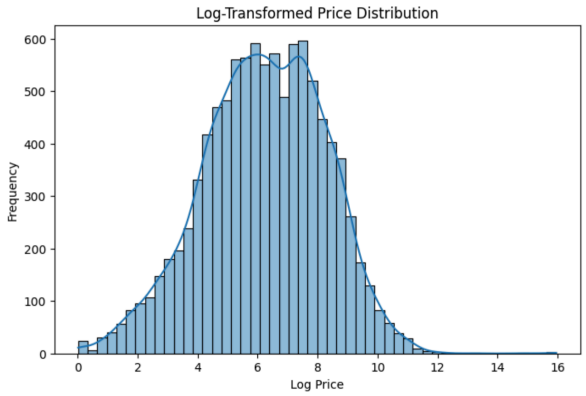
Linear regression was the weakest performer of classical models, which likely stems from the models' inability to account for the complex, nonlinear relationships inherent when it comes to pricing data, as well as its sensitivity to multicollinearity and outliers. And finally, the neural network had the poorest test performance, despite success with training on preprocessed features. R^2 = -4.0233, MSE = 19.4413, and MAE of 3.9900 suggests that substantial overfitting or a misalignment between network architecture and feature structure. While the model was able to minimize loss during training, it was unable to effectively generalize unseen data. These findings reinforce the importance of regularization techniques and more rigorous tuning when using neural networks on tabular data with high feature sparsity.

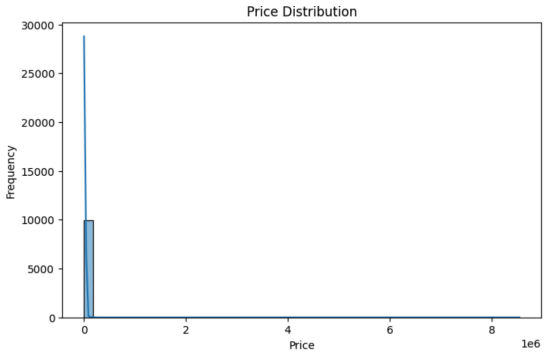
Ultimately, the results show how tree-based models such as Random Forests are vital for healthcare pricing prediction tasks especially as far as interpretability, feature interactions, and robustness to mixed data types are important.

Exploratory Data Analysis

The goal of the exploratory data analysis is to develop a better understanding of the hospital pricing dataset, identify key trends, assess data quality, and prepare the foundation for predictive modeling. The dataset that's derived from Sage Transparency's published hospital billing data is used in this EDA to visualize data distributions, detect anomalies, and assess relationships between variables that may influence hospital service pricing.

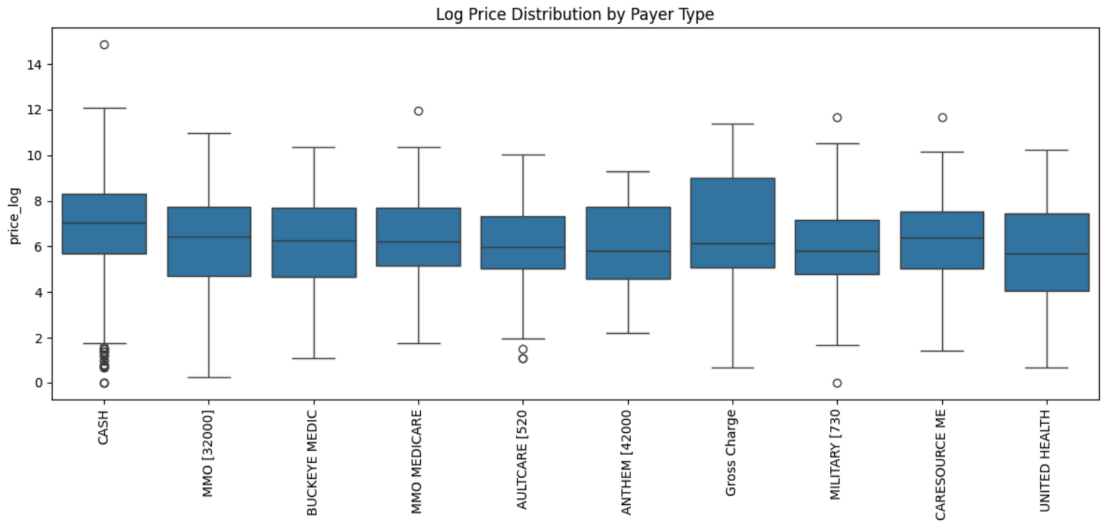
To visualize the skewness of the data, we plotted a histogram of the raw price variable. The histogram confirmed a classic long right tail, suggesting the presence of many low and mid range prices alongside a few extreme outliers, which is a common phenomenon in healthcare billing data. To address the skew, we applied a logarithmic transformation to the price variable, with the resulting distribution of log(price) approximating a normal distribution, which is advantageous for linear and neural network regression models. Log transformation also helps to stabilize variance and reduce the impact of outliers on model training.



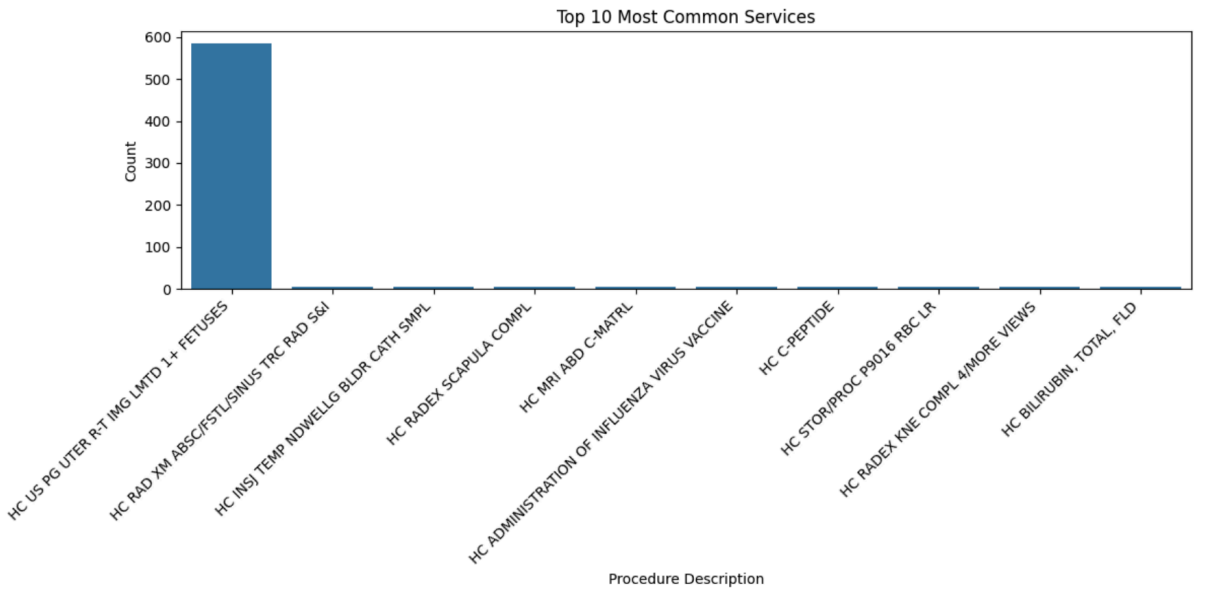


It's also crucial to understand how prices vary across payer types. We generated boxplots of prices grouped by payer, meaning involving Medicare, Medicaid, Commercial, and Self Pay. Notable observations include the fact that Medicare and Medicaid tend to reimburse at the lowest and most consistent rates, Commercial payers show substantial variability with significantly

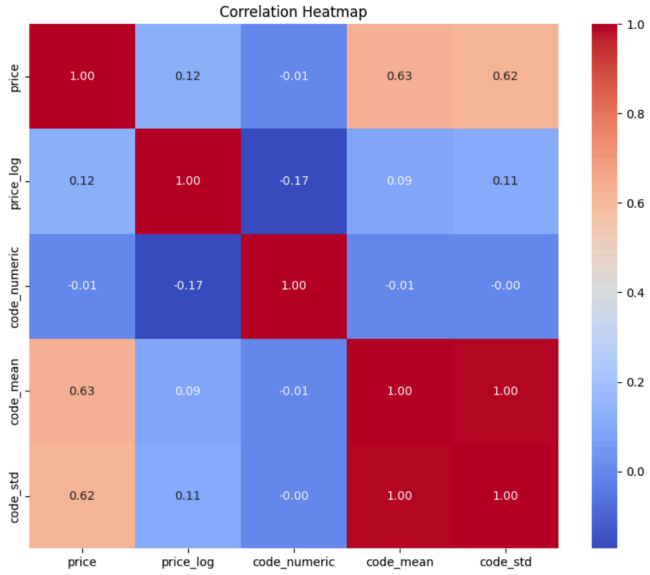
higher prices, and Self-pay prices are generally higher than Medicare/Medicaid but also with higher variance. These observations support the hypothesis that payer type is a strong determinant of pricing variation, and reflect the different negotiation powers and policy constraints among payer groups.



To gain insight into what services are most frequently recorded, we also created a bar chart of the top 10 most common services in the dataset. These included basic metabolic panels, CT scans, colonoscopies, mammograms, and emergency room visits. Each of these services appear thousands of times in the dataset, providing sufficient data to support robust model training. Further, the pricing varies widely depending on payer and hospital.



We also computed Pearson correlation coefficients between numerical variables, with a correlation heatmap produced to visualize these relationships. The key variables consist of price (raw and log transformed), price_percent_of_medicare, net_patient_revenue, and discharges. Key insights that can be found from this analysis include the observation of a moderate positive correlation between price and price_percent_of_medicare, indicating that hospitals charging more often do so across payers. There is also a weak correlation between price and net patient revenue or discharges, suggesting that hospital scale has only limited explanatory power for price variation.



The key takeaways from the EDA show that price is a skewed variable with extreme outliers. Log transformations are necessary for appropriate modeling. Payer type is a critical determinant of price variation, meaning it should be treated as a key feature, while service type and market geography also exhibit clear associations with price. Meanwhile, high cardinality categorical features, such as system_name, require encoding strategies to avoid sparsity. And finally, data cleaning and transformation methods play a central role in enabling accurate and interpretable modeling. The EDA provides a strong foundation for developing predictive models that can capture the complex relationships between inherent hospital pricing, and these insights can guide feature selection, model choice, and evaluation metrics going forward.

Future Directions

Given more time, there are many extensions and improvements to be had on this project. First, the neural network could benefit from additional hyperparameter tuning, architectural experimentation, and regularization strategies to improve generalization. Incorporating a validation split or early stopping could possibly mitigate overfitting, which was a prominent issue in this implementation.

Feature engineering could also be expanded to incorporate interaction terms or embeddings for high cardinality categorical variables such as hospital system or service codes. Future work could also explore temporal dynamics by incorporating publish dates or longitudinal changes in hospital pricing data.

Data wise, external datasets such with socioeconomic indicators, patient outcomes, and so forth could further enrich feature space and support more nuanced modeling. In line with that, the code could also be refactored to be able to handle the full dataset without sampling. Comparative studies could benchmark gradient boosting or hybrid models against the current best performer. These steps could address current limitations while amplifying the impact for healthcare stakeholders.

Conclusion

This project set out to find the effectiveness of various machine learning models in predicting hospital procedure prices using data that's publicly available. Several predictive models were built and assessed, consisting of Random Forest, K_nearest Neighbors, Linear Regression, and a neural network, using a dataset with many features at the hospital, service, and market levels. The Random Forest model significantly outperformed the others, with the highest accuracy and demonstrating a strong ability to capture nonlinearities and contextual pricing patterns.

The neural network architecture showed promise during training, but failed when it came to generalization, and traditional models such as linear regression underperformed, which reaffirmed the understanding that simple assumptions are typically not adequate when dealing with complex, high variance healthcare pricing data.

This work highlights the importance of robust preprocessing, thoughtful feature engineering, and systematic model comparison in applied machine learning. It also contributes to the broader healthcare transparency movement by showcasing how algorithmic methods can be used to reveal and predict pricing behavior across institutions and payers.

Ultimately, this project demonstrates how data science can be used for predictive modeling as well as supporting greater accountability in healthcare systems. With the increase of regulatory mandates increasing transparency requirements, predictive tools like the ones explored here have the potential to inform consumers, policy makers, and providers alike.

References

Bai, G., & Anderson, G. F. (2015, June). Extreme markup: The Fifty US hospitals with the highest charge-to-cost ratios. PubMed. https://pubmed.ncbi.nlm.nih.gov/26056196/

Bertsimas, D., Bjarnadóttir, M. V., & Kane, M. A. (2008, November). Algorithmic prediction of health-care costs. PubsOnLine. https://pubsonline.informs.org/doi/pdf/10.1287/opre.1080.0619

Huang, A. W., Haslberger, M., Coulibaly, N., Galárraga, O., Oganisian, A., Belbasis, L., & Panagiotou, O. A. (2022, March). Multivariable prediction models for health care spending using Machine Learning: A protocol of a systematic review. Diagnostic and prognostic research. https://pubmed.ncbi.nlm.nih.gov/35321760/

Huang, Y., Talwar, A., Chatterjee, S., & Aparasu, R. R. (2021, May 6). Application of machine learning in predicting hospital readmissions: A scoping review of the literature. BMC medical research methodology. https://pubmed.ncbi.nlm.nih.gov/33952192/

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., … Dean, J. (2018, May 8). Scalable and accurate deep learning with electronic health records. Nature News. https://www.nature.com/articles/s41746-018-0029-1