Deep Learning Model for maximizing hospital equipment sale, utilizing propensity scores

Group Members: Jennifer Nasirumbi, Udayveer Singh Andotra, Andrew Walker

Course: 960:588 Data Mining, Fall 2024

Instructor: Professor Javier Cabrera

Summary

This project aimed to identify potential new customers for orthopedic products by analyzing over four thousand hospitals across the United States. Using a combination of SuperLearner ensemble methods and deep learning models on transformed data, we accurately projected sales for the top ten selected hospitals, which matched or exceeded the sales of current top customers. Approximately half of these high-potential hospitals were located in the Northeast, and two were in the Northwest, aligning with regions of existing high sales. We focused on hospitals with a propensity score between 0.35 and 0.65 that were not current customers. Factor Analysis confirmed the possibility of reducing data dimensions, though all relevant features. The integrated approach of SuperLearner and deep learning proved effective in pinpointing the most promising hospitals to target for the sales campaign, optimizing our strategy for maximum impact.

1. Introduction

This study aims to boost sales of our orthopedic products to hospitals across the United States. By identifying non-customer hospitals similar to our existing customers, we can effectively target new potential clients. Using the SuperLearner algorithm, we estimate propensity scores to pinpoint these hospitals. Following this, a deep learning model predicts potential sales for the identified hospitals, allowing us to focus on those with the highest sales potential. This targeted approach aims to maximize our sales efforts and expand our customer base.

2. Data Overview

The dataset used in this study, referred to as **hospitalUSA**, contains various attributes related to hospitals across the United States. Below is a detailed description of the variables included in the dataset:

- **ZIP**: US Postal Code. Identifies the geographic location of each hospital.
- **HID**: Hospital ID. A unique identifier assigned to each hospital.
- **CITY**: City Name. The city where the hospital is located.
- **STATE**: State Name. The state where the hospital is located.
- **BEDS**: Number of Hospital Beds. Indicates the capacity of the hospital in terms of available beds.
- **RBEDS**: Number of Rehab Beds. Indicates the capacity of the hospital specifically for rehabilitation patients.
- **OUT-V**: Number of Outpatient Visits. The total number of outpatient visits recorded annually.
- **ADM**: Administrative Cost. Annual administrative costs in $1000s.
- **SIR**: Revenue from Inpatient. Annual revenue from inpatient services in $1000s.
- **SALES**: Sales of Rehab Equipment. Annual sales of rehabilitation equipment in $1000s.
- **HIP**: Number of Hip Operations. Annual number of hip operations performed.
- **KNEE**: Number of Knee Operations. Annual number of knee operations performed.
- **TH**: Teaching Hospital. Binary indicator (0 or 1) indicating whether the hospital is a teaching hospital.
- **TRAUMA**: Trauma Unit. Binary indicator (0 or 1) indicating whether the hospital has a trauma unit.
- **REHAB**: Rehab Unit. Binary indicator (0 or 1) indicating whether the hospital has a rehabilitation unit.
- **HIP2**: Number of Hip Operations (Year 2). Annual number of hip operations performed in the second year.
- **KNEE2**: Number of Knee Operations (Year 2). Annual number of knee operations performed in the second year.
- **FEMUR2**: Number of Femur Operations (Year 2). Annual number of femur operations performed in the second year.

3. Methodology

- **Data Preparation**:

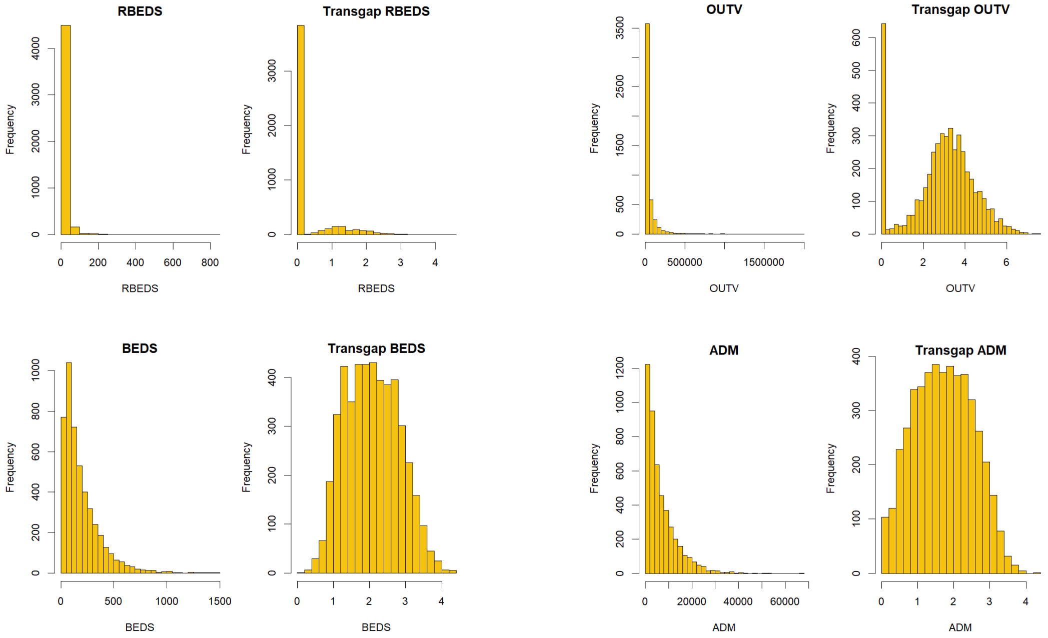
Definition of group variable C: The group variable C is used to differentiate between hospitals that are current customers and those that are not. This binary variable plays a crucial role in the propensity score estimation process.

Value 1: Assigned to hospitals where **Sales > 0**, indicating that these hospitals are already purchasing our orthopedic products.

Value 0: Assigned to hospitals where **Sales = 0**, indicating that these hospitals are not currently our customers.

Data cleaning and transformation: The `transgap` function is designed to find optimal transformations for the feature variables to improve the performance of predictive models. It enhances the data by addressing issues such as skewness and gaps. Reasons for using transgap -

- Applies various power and logarithmic transformations to the data.
- Selects the transformation that minimizes skewness and gaps, ensuring a more normal distribution of the data.
- Handles negative values by shifting and changing signs if necessary.



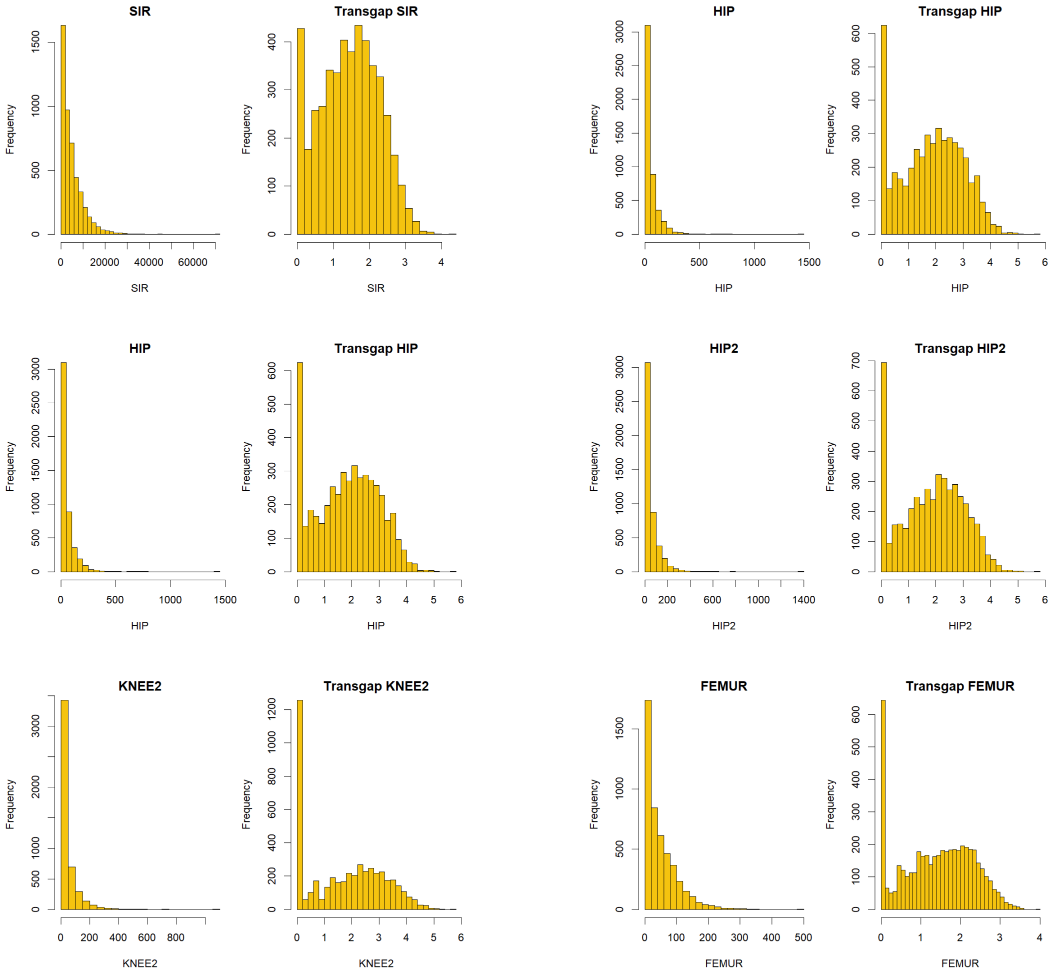


Figure1: Continuous variables and their transformations

- ## Propensity Score Estimation

  To estimate propensity scores, we first prepare the data by creating a group variable C to identify customer and non-customer hospitals, where C=1 for customers and C=0 for non-customers. We fit a SuperLearner model using an ensemble of algorithms, including generalized linear models, random forests, support vector machines, and neural networks. This model estimates the propensity scores, which indicate the likelihood of each hospital being similar to our current customers. We then evaluate the model's performance using cross-validation and a confusion matrix to ensure its accuracy.

  Finally, we use the propensity scores to identify a subset of non-customer hospitals that are most similar to our current customers, focusing on those with scores between 0.35 and 0.65 to optimize our sales targeting efforts. Hospitals with very high scores might already be well-served by our competitors or have reached their capacity for purchasing similar products. This may limit the potential for significant sales growth. Also there's a possibility of encountering diminishing returns as these hospitals might already be utilizing similar orthopedic products extensively. So we think a range of 0.35 to 0.65 is optimal.

Prediction	Reference		
		0	1
	0	1979	14
	1	17	2693

Accuracy : 0.9934

95% CI : (0.9907, 0.9955)

Table1: Performance statistics

Sensitivity	0.9915
Specificity	0.9948
Pos Pred Value	0.9930
Neg Pred Value	0.9937
Prevalence	0.4244
Detection Rate	0.4208
Detection Prevalence	0.4238
Balanced Accuracy	0.9932
Positive Class	0

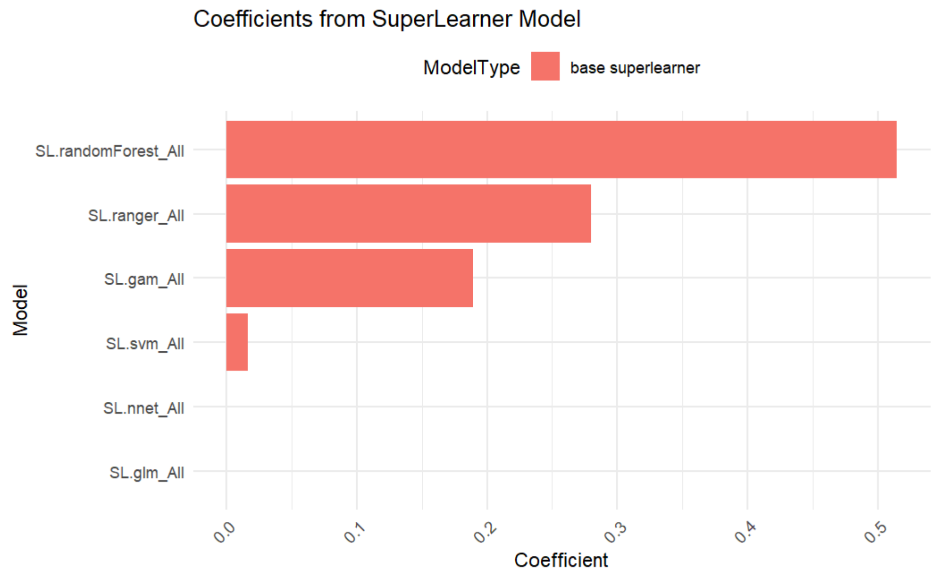


Figure2: Super Learner Coefficients after CV

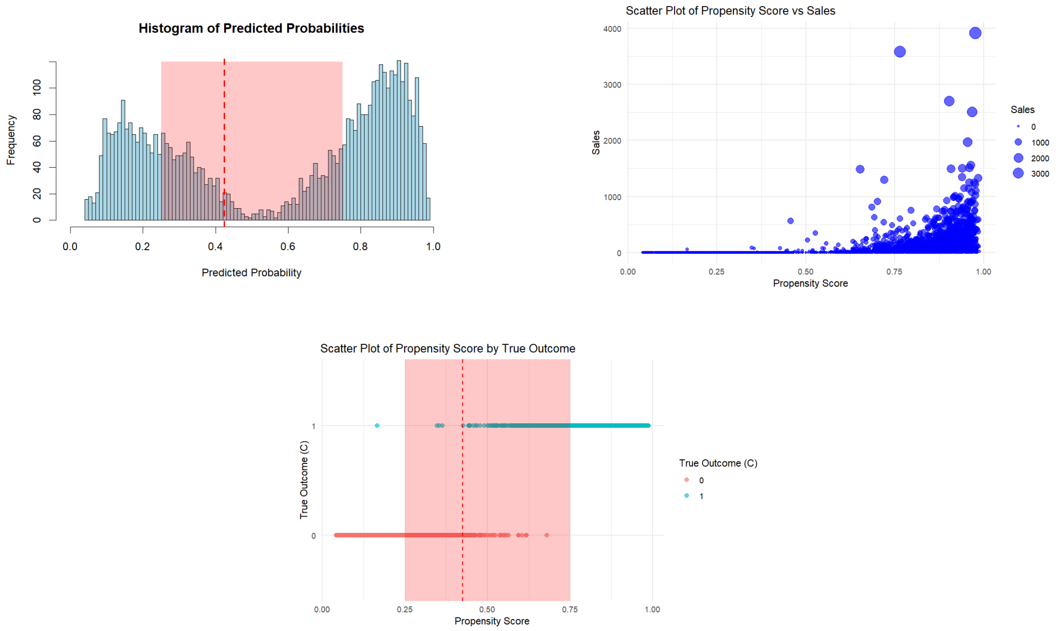


Figure3: Distribution Propensity scores and the true outcomes

5. Predictive Modeling

To predict potential sales for the identified subset of non-customer hospitals, we used h2o to build a deep learning model, specifically a feedforward neural network. This type of neural network is ideal for regression tasks. The model was constructed with an input layer to accommodate the hospital features, followed by two hidden layers with ReLU activation functions to capture complex patterns, and a final output layer with a linear activation function to predict continuous sales values. The model was compiled with the Adam optimizer and Mean Squared Error (MSE) loss function, then trained on the data. After training, the model predicted sales for the targeted subset, which was then evaluated using metrics like MSE and R-squared. This approach enabled us to identify the top 10 hospitals with the highest predicted sales potential, providing a data-driven strategy to optimize sales efforts.

Variable Selection

Factor analysis is used to uncover the underlying relationships between variables by reducing the dimensionality of the data. This technique helps in identifying the core factors that drive variations in the data, which can then be used as inputs in the deep learning model to enhance its performance.

Variable importance helps to identify which features significantly influence the outcome. This is crucial for model interpretation and improving predictive performance by selecting the most relevant features.

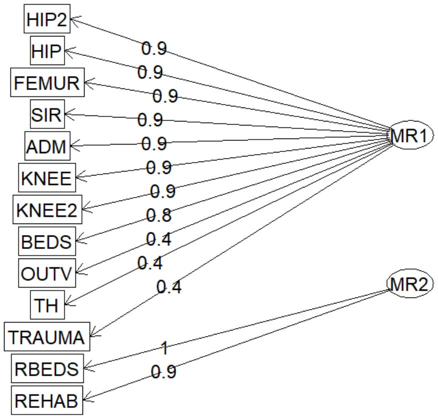
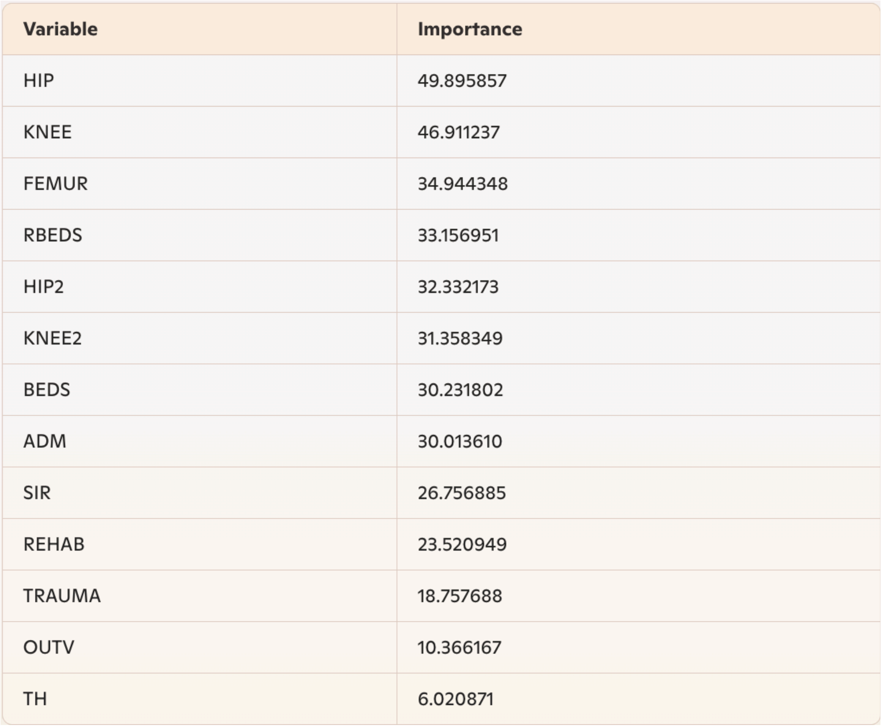


Figure4: Factor Analysis

In this study, we use the Random Forest model to compute variable importance. The Random Forest algorithm evaluates the importance of each variable by measuring the increase in prediction error when the values of that variable are permuted while all others are left unchanged. Variables that greatly increase the error are considered more important.

Table3: Variable Importance



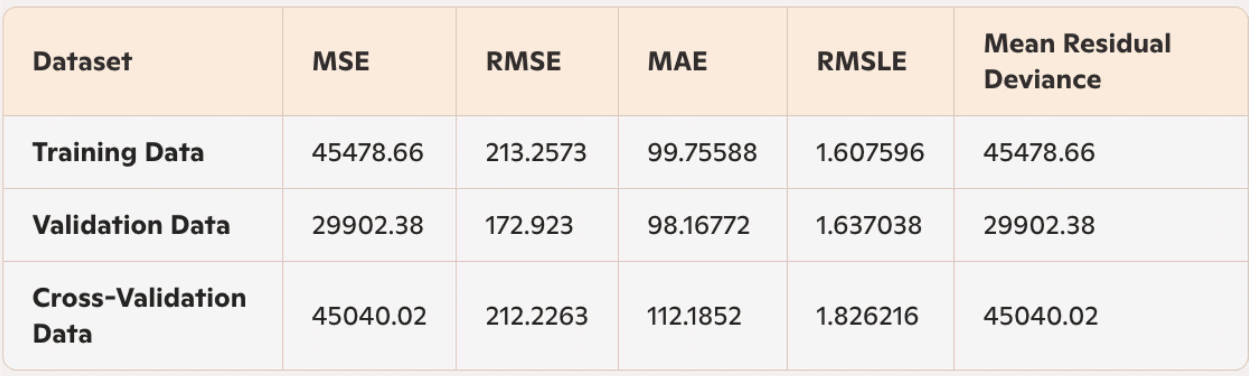
Variable	Importance
HIP	49.895857
KNEE	46.911237
FEMUR	34.944348
RBEDS	33.156951
HIP2	32.332173
KNEE2	31.358349
BEDS	30.231802
ADM	30.013610
SIR	26.756885
REHAB	23.520949
TRAUMA	18.757688
OUTV	10.366167
TH	6.020871

Despite the factor analysis indicating that the data could be reduced to two dimensions, we chose to use all the variables listed above in the deep learning model. This approach ensures that we leverage all relevant information, excluding only the non-informative categorical variables such as “state,” “city,” “hospital id,” and “zip code,” to enhance the model's predictive accuracy and robustness.

Model Overview:

- **Type**: H2ORegressionModel: deeplearning
- **Model Key**: DeepLearning_model_R_1734810548184_26
- **Layers**: Regression with Gaussian distribution and quadratic loss
- **Weights/Biases**: 3,521
- **Model Size**: 48.6 KB
- **Training Samples**: 123,751
- **Mini-Batch Size**: 1

Table3: Performance Metrics



Dataset	MSE	RMSE	MAE	RMSLE	Mean Residual Deviance
Training Data	45478.66	213.2573	99.75588	1.607596	45478.66
Validation Data	29902.38	172.923	98.16772	1.637038	29902.38
Cross-Validation Data	45040.02	212.2263	112.1852	1.826216	45040.02

6. Results

The table 4 below lists the top 10 hospitals based on their predicted sales potential. The HID column represents the unique hospital identifier, while the ZIP, CITY, and STATE columns provide location information. The Predicted Sales column shows the estimated sales value for each hospital, derived from our deep learning model. The Potential Gain column indicates

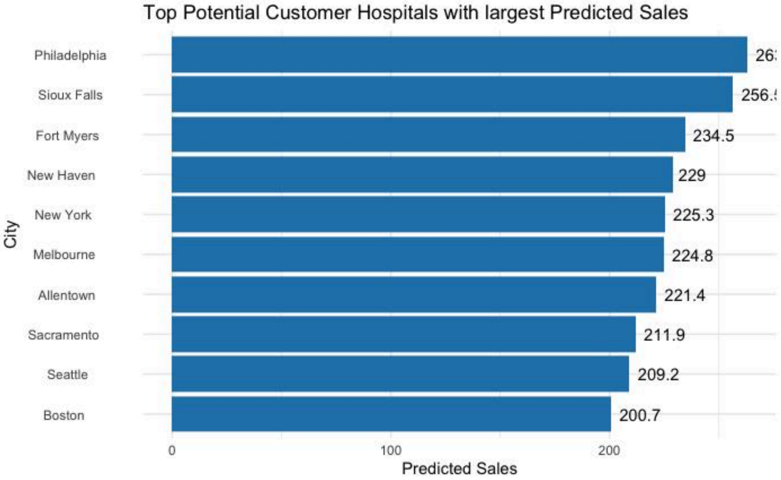
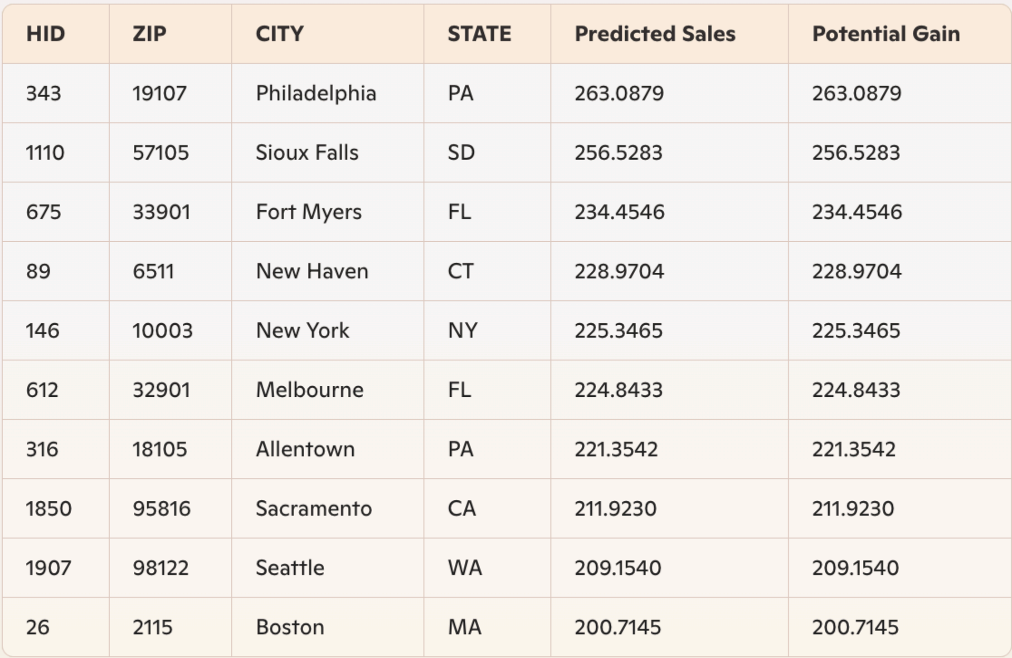


Figure5: Table4 graph

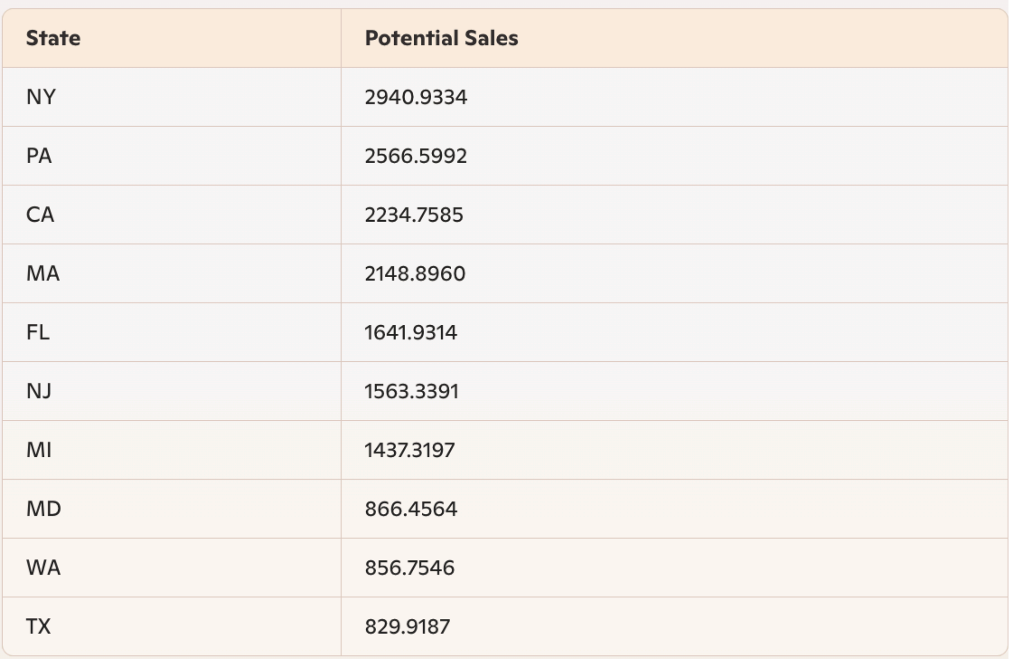
the same value, highlighting the potential revenue increase if these hospitals were to be targeted for sales efforts. The potential sales for by city(Table4) and by state is in 1000s of $.

Table4



HID	ZIP	CITY	STATE	Predicted Sales	Potential Gain
343	19107	Philadelphia	PA	263.0879	263.0879
1110	57105	Sioux Falls	SD	256.5283	256.5283
675	33901	Fort Myers	FL	234.4546	234.4546
89	6511	New Haven	CT	228.9704	228.9704
146	10003	New York	NY	225.3465	225.3465
612	32901	Melbourne	FL	224.8433	224.8433
316	18105	Allentown	PA	221.3542	221.3542
1850	95816	Sacramento	CA	211.9230	211.9230
1907	98122	Seattle	WA	209.1540	209.1540
26	2115	Boston	MA	200.7145	200.7145

Table5: Potential Sales to Current Non-Customers by State



State	Potential Sales
NY	2940.9334
PA	2566.5992
CA	2234.7585
MA	2148.8960
FL	1641.9314
NJ	1563.3391
MI	1437.3197
MD	866.4564
WA	856.7546
TX	829.9187

7. Conclusion

From the analysis of over four thousand hospitals that was performed to identify hospitals that were likely candidates to target with the Sales campaign. It was observed that using SuperLearner ensemble Followed by deep learning on the pre-transformed data of the top ten potential hospitals selected, the projected sales were similar or even better than the current top customers. Of the 10 hospitals that were recommended, about half of them were hospitals in the NorthEast while two were from the NorthWest. This is consistent with the current customers with highest sales. We chose hospitals with a propensity score between 0.35 and 0.65 that were not current customers. Factor Analysis proves that the data could be reduced to two dimensions as shown. We conclude that amalgamation of the 2 methods, the superlearner and deep learner were effective in identifying the most likely hospitals to target in the sales campaign.

8. Appendix

R code file and datasets used can be found in the zipped file.