Genesis 10 – Dev10 Data Track

Public

Capstone Project Technical Report

Alysia Halverson, Charles Rehder, Jakob Thunen
5-26-2022

Introduction

This capstone project will use a variety of data literacy, data analysis, data engineering, and data science concepts and techniques to research, better understand, and predict payments regarding public health care. Our research focuses on Medicare and Medicaid Services and examines several measures of quality of care, payment prices, and demographics of U.S. citizens enrolled in these public health insurance systems. Some of the tools and concepts used in this project include Python (`pandas`, `scikit-learn` and `pyspark`), Machine Learning, Azure products, and Kafka (Confluent).

Most of the data used to better understand quality of care originates from the Hospital Value-Based Purchasing (HVBP) Program which links Medicare's payment system to health care quality in the inpatient setting, a long-standing effort by the Centers for Medicare and Medicaid Services (CMS). CMS utilizes the first ever national pay-for-performance program by evaluating hospitals across several value of care domains to provide payments to hospitals based on the quality of care, rather than only the quantity of services provided. For this project, data from four of these domains will be utilized to research, analyze, and make predictions. In addition, data from CMS regarding payment and value of care provides information regarding payment for the same hospitals including comparisons between actual payments and the Medicare Spending Per Beneficiary (MSPB), hospitalization reasoning, and value of care results for actual mortality and complication measures. Finally, data from the U.S. Census Bureau will provide contextual information on residents enrolled in public health insurance programs, specifically Medicare, Medicaid, and Veterans Affairs (VA), as well as population counts.

Using machine learning, this project's primary purpose is to accurately predict payment prices based on features from the data. Payment price, terms used interchangeably within this report and report's materials, is referring to payments made by Medicare patients or on behalf of Medicare patients for health care services starting on the first day of a hospitalization through the next 30 or 90 days, depending on the hospitalization reason, for Medicare fee-for-service beneficiaries. Hospitalization reasons include heart attack, heart failure, pneumonia, and total hip/knee replacements. These payments, also known as risk-standardized payments (RSPs), are consistent with the priorities of the framework of the Department of Health and Human Services to ensure effective communication, coordination of care, treatment, community engagement, affordable care, harm reduction, and collaboration between Medicare, hospitals, and government agencies. Because of the priority set by CMS with the HVBP, there are many associated features that we suspect to impact payment: clinical outcomes, safety, personal and community engagement, and efficiency. Our goal is to be able to reliably predict prices with these values by trying several different regression models. R^2 scores will be given for our linear regression payment model, lasso regression payment model, random forest regressor model, and finally, our XGBoost model. After tuning each model using the proper hyperparameters, we found that the XGBoost model had the highest accuracy score and was able to predict price most reliably with numerous given values in the previously mentioned HVBP program domains. All code, project specifications, and deliverables can be in this project's Github (linked).

Exploratory Questions

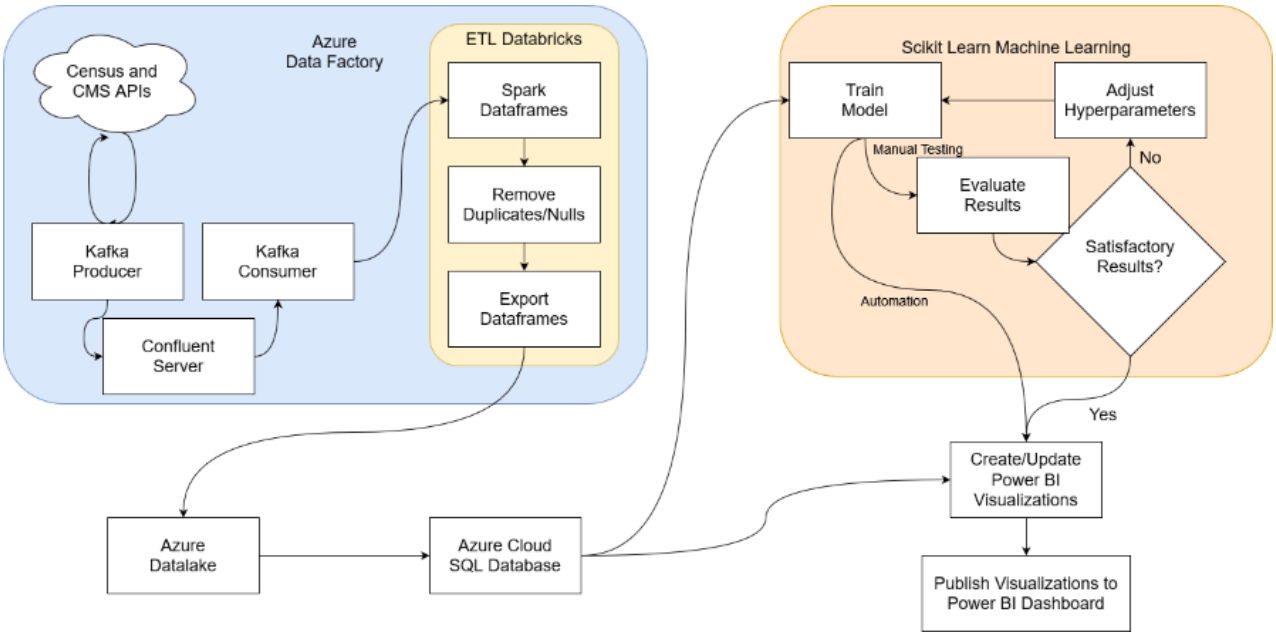
As previously discussed, this project has a few purposes and goals along with the utilization of machine learning to predict price of hospitalizations. This project will also seek to better determine which measured factors have the greatest impact on patient care quality and identify possible trends or patterns across different states using demographic information from the Census Bureau. To lead our efforts, we identified several questions we sought to answer:

Do price and value of care always align?
Will safety have a large influence on the price of hospitalizations?
Would higher efficiency scores decrease the price of hospitalizations or increase them?
How might community engagement influence the price of hospitalizations?
How might clinical outcomes influence the price of hospitalizations?
What are trends among U.S. states and those insured on public healthcare?
Are there patterns between value of care and states with higher public insurance rates?
Do hospitals with higher safety scores suffer with efficiency scores?
Do states that are more/less urbanized have higher/lower community engagement scores?
Do hospitals with higher community engagement scores have higher clinical outcomes scores?

Data Sources and Information

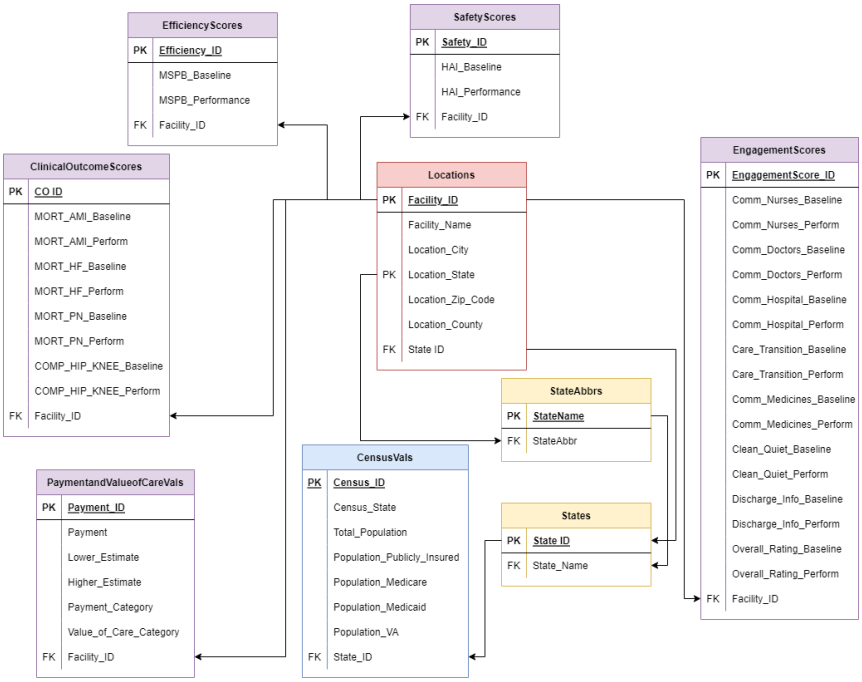
We utilized Azure Data Factory to extract, clean, transform, and export our data. Using Azure Databricks, we called the API for both data sources and created two Kafka Producer streams to simulate live, streaming data from both sources. Using created topics, the messages from the Confluent server are consumed by two Kafka Consumers and the CMS data is migrated from Kafka into an Azure Data Lake as CSV files. ETL is then completed using `pyspark` within individual databricks for each of the five CMS datasets. For more detailed information, please see our repeatable ETL report located on our Github. The data was then written into the Azure SQL database and can be used for machine learning, visualizations, and Power BI.

DATA PROCESS: CLOUD ETL, MACHINE LEARNING, & POWER BI DASHBOARD REPORT



For use in both machine learning model building and visualizations, a static environment was created and the database was loaded in bulk using the above process, outside of Azure Data Factory automation, on May 13, 2022. We created a temporary database for automation so as not to disrupt the work of machine learning and creating the Power BI dashboard report.

ERD FOR DATABASE IN AZURE SQL DATABASE (GEN10 SERVER)



THE CENTERS FOR MEDICARE AND MEDICAID SERVICES (CMS)

To better understand health insurance quality of care, specifically related to Medicare and Medicaid payments and quality of care measures, we utilized the following five datasets. A complete data dictionary can be found here (linked). All the following datasets have a related variable, Facility ID, that we use to merge the data with joins.

- *Hospital Value-Based Purchasing (HVBP) - Efficiency Scores*. (2022, April 27). [Dataset]. https://data.cms.gov/provider-data/dataset/su9h-3pvj
    - The Efficiency and Cost Reduction dataset provides details on the Medicare Spending per Beneficiary (MSPB) measure. The MSPB measure is the hospital's MSPB measure ratio calculated as MSPB amount (hospital's Medicare spending per beneficiary dollar amount) divided by the median MSPB amount (national median MSPB dollar amount).
- *Hospital Value-Based Purchasing (HVBP) - Person and Community Engagement Domain Scores (HCAHPS)*. (2022, April 27). [Dataset]. https://data.cms.gov/provider-data/dataset/avtz-f2ge
    - The Person and Community Engagement Domain Scores dataset provides details on eight dimensions related to patient experience of care collected from Hospital

Consumer Assessment of Healthcare Providers and Systems (HCAHPS) surveys. CMS calculates these scores using the "top-box raw score", which is the unrounded percentage of a hospital's patients who chose the most positive response to the HCAHPS survey items. More information about these scores can be found in the HVBP Step-by-Step Guide (linked). Higher percentages correspond to higher quality.

- *Hospital Value-Based Purchasing (HVBP) - Safety*. (2022, April 27). [Dataset]. https://data.cms.gov/provider-data/dataset/dgmq-aat3
    - The Safety Measures dataset provides details on healthcare-associated infections (HAI). These measures are standardized infection ratios (SIRs) using the number of observed infections (numerator) divided by the number of predicted infections, calculated by the CDC, (denominator). Lower values correspond to higher quality.
- *Hospital Value-Based Purchasing (HVBP) - Clinical Outcomes Domain Scores*. (2022, April 27). [Dataset]. https://data.cms.gov/provider-data/dataset/pudb-wetr
    - The Clinical Outcomes dataset provides details on mortality on Acute Myocardial Infarction (AMI), Heart Failure (HF), Pneumonia (PN), and Total Hip/Knee Arthroplasty (COMP-HIP-KNEE) in a survival ratio. Higher values indicate better outcomes.
- *Payment and value of care - Hospital*. (2022, January 26). [Dataset]. https://data.cms.gov/provider-data/dataset/c7us-v4mf
    - The Payment and Value of Care dataset provides details on payments and MSPB comparisons. The dataset also includes the payment category (*Is the payment greater, no different, or less than the average payment?)* and the value of care category (*Is the mortality better, average, or worse? Are the complications average or worse? Was the payment high, low, or average?*). Our ML models will attempt to predict payment price using these features, mapped to be numeric, and the previously mentioned features using HVBP domain scores.

UNITED STATES CENSUS BUREAU

- U.S. Census Bureau, *2015-2019 5-Year American Community* Survey, Table S2704: Public Health Insurance Coverage by Type and Selected Characteristics (All States within United States). (2020) [Data set]. https://data.census.gov/cedsci/table?t=Health%20Insurance&g=0100000US%240400000&tid=ACSST5Y2020.S2704&tp=true
    - The Public Health Insurance Coverage by Type and Selected Characteristics (All States within United States) dataset will provide population counts for the state, the number of residents on public health insurance alone, and the breakdown populations for each type of public health insurance: Medicare, Medicaid, and VA-only. This dataset is joined using the State/State ID with our other data.

Research Process, Findings, and Visualizations

CENSUS BUREAU: PUBLIC HEALTH INSURANCE & STATES

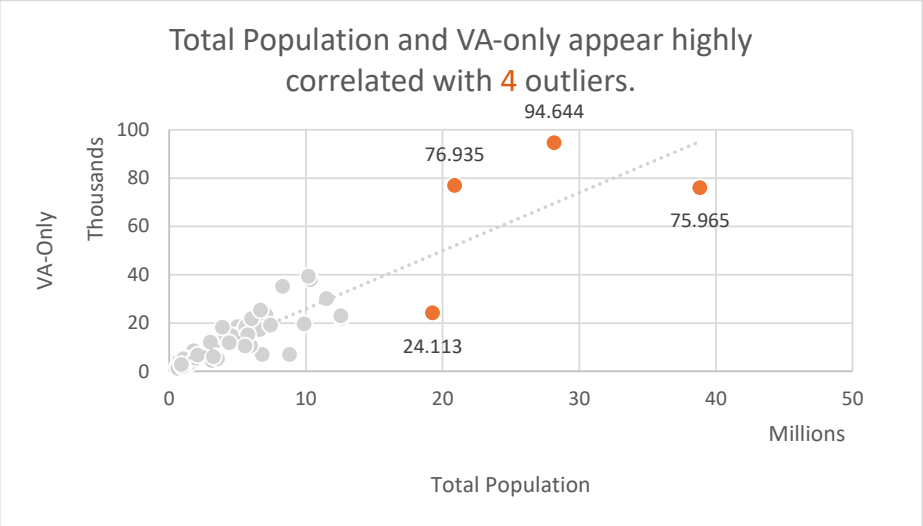
What are trends among U.S. states and those insured on public healthcare?

First, to better understand the demographics by state regarding residents insured in public health insurance programs, we completed some initial descriptive statistics and visualizations to identify possible trends or patterns. For this, we used the previously mentioned table S2704 from the U.S. Census Bureau which provides information by type of public health insurance and total population counts as well as population counts for each enrollment type: Medicare, Medicaid, and VA. When running a correlation matrix on states' total population and publicly insured population, we found a strong positive correlation between the two.

	Total Population
Total Population	1
Publicly Insured Population	**0.981**

The number of residents under public health insurance alone is strongly correlated with the total state population outside of a few outliers. California, the most populated state with 38.8 million residents, appears as an outlier due to a higher count of residents enrolled in public health insurance alone, totaling 9.7 million residents. Texas, on the other hand, appears as an outlier due to a total population of 28.1 million residents and just half of the number of residents enrolled in public health insurance alone compared to California, totaling 4.9 million residents. The remaining two outliers, Florida and New York, have notably less distinction but New York falling slightly above the trendline, meaning more residents enrolled in public health insurance, and Florida falling slightly below the trendline, meaning less residents enrolled in public health insurance.

Interestingly, these four states remain outliers when comparing correlation between total population and VA-only insured but on opposite spectrums; California and New York falling below the trendline, meaning less residents enrolled in VA-only public health insurance, and Texas and Florida falling above the trendline, meaning more residents enrolled in VA-only public health insurance.



This suggests that when considering VA programming, advertising, and planning, knowledge about states with significantly higher proportions of VA-only health insurance like Texas and Florida is key. This may also suggest to agencies like CMS or hospital administrators to take these populations into consideration when estimating cost of hospitalizations or communicating with those insured by the VA.

Are there patterns between value of care and states with higher public insurance rates?

As a first step of data analysis, we decided to run a correlation matrix across value of care scores and public insurance population counts among states to see if there were any obvious relationships. From this analysis, it is apparent that there is a strong positive relationship between states' population of residents enrolled in public health insurance and count of reported hospitalizations, meaning reporting is a seemingly accurate sample size. Outside of this and a low-to-moderate positive correlation between average payment amount and states' publicly insured population, there is low to no correlation found between value of care metrics in engagement scores, clinical outcomes (complication rate and survival rate), safety ratio, or efficiency (MSPB) ratio and states with higher public insurance rates.

	Total Population	***Publicly Insured Population***
Total Population	1	
Publicly Insured Population	0.983	1
Average Payment	0.477	0.444
Overall Engagement Score Rating	-0.177	**-0.195**
Count of Reported Hospitalizations	0.955	0.916
Average Complication Rate of Total Hip/Knee Arthroplasty	-0.058	**-0.055**
Average Survival Rate of AMI, HF, and PN	0.262	**0.255**
Average Safety Ratio	0.100	**0.100**
Average Efficiency Ratio	0.186	**0.174**

Do states that are more/less urbanized have higher/lower community engagement scores?

When looking at the states with the lowest overall community engagement scores, five of the top ten urbanized states rank in the bottom ten: New York, Florida, New Jersey, Georgia, and Michigan. Out of the most urbanized states, Texas ranked the highest at #16 with 71.6%. When running data analysis on the correlation between percentage of urban population and average of overall community engagement scores, there is a very small negative correlation. Thus, we conclude that these variables are mostly unrelated to one another.

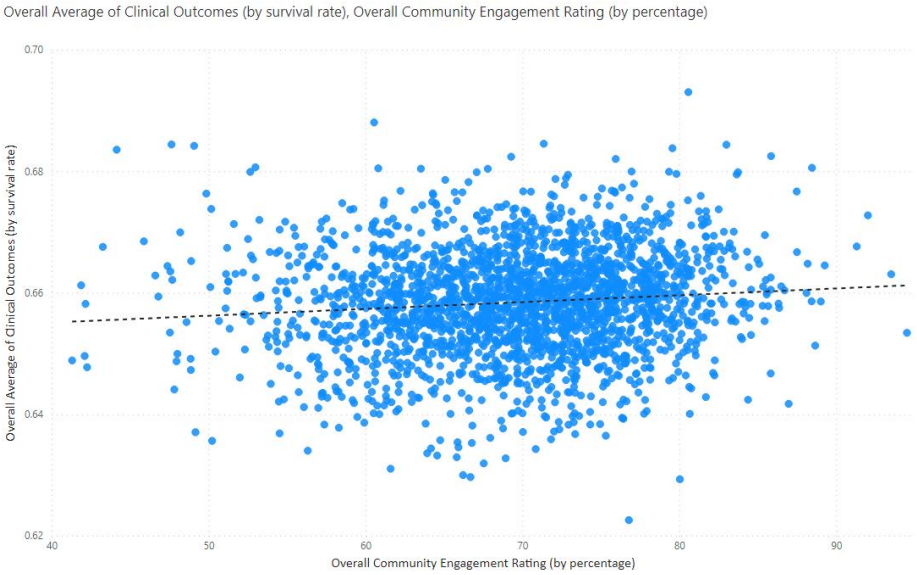
	Publicly Insured Residents	***Percentage of Urban Population***
Publicly Insured Residents	1	
Percentage of Urban Population	0.381	1
Average of Overall Rating Of Hospital Performance Rate	-0.195	**-0.214**

HVBP – Quality of Care (Efficiency, Safety, Person/Community Engagement, & Clinical Outcomes)

As previously stated, the purpose of the HVBP program is to assess quality of care with the assumption that positive scores in each of the domains lead to better clinical outcomes and therefore more efficient Medicare and hospital spending. As such, many of our exploratory questions were guided by this assumption and our research will hope to answer some of these relationship-based questions and investigate possible correlations between the HVBP program domains, payment, value of care, and outcomes.

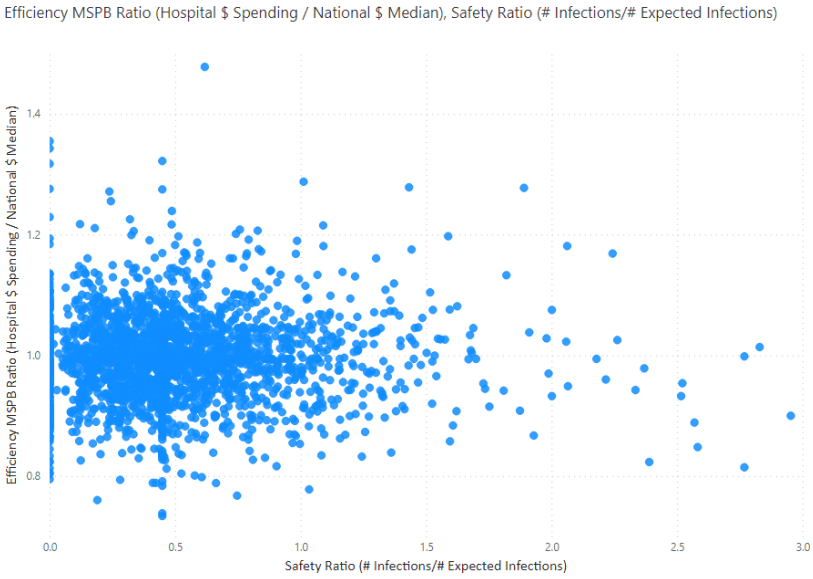
Do hospitals with higher community engagement scores have higher clinical outcomes scores?

When determining the relationship between community engagement scores and higher clinical outcomes, the visualization below shows the very small correlation between the two. Therefore, we can conclude that community engagement scores alone do not greatly impact clinical outcomes.



Do hospitals with higher safety scores suffer with efficiency scores?

When determining the relationship between hospital safety scores and efficiency scores, the below visualization shows the lack of a correlation between the two. Therefore, we can conclude that safety scores alone do not greatly impact efficiency scores. When looking at this data, it is important to note that the data source for safety was highly impacted by the pandemic and had mostly null values. To better assess this potential relationship between safety scores and efficiency scores, it would be beneficial to conduct this similar data analysis in a non-pandemic year.



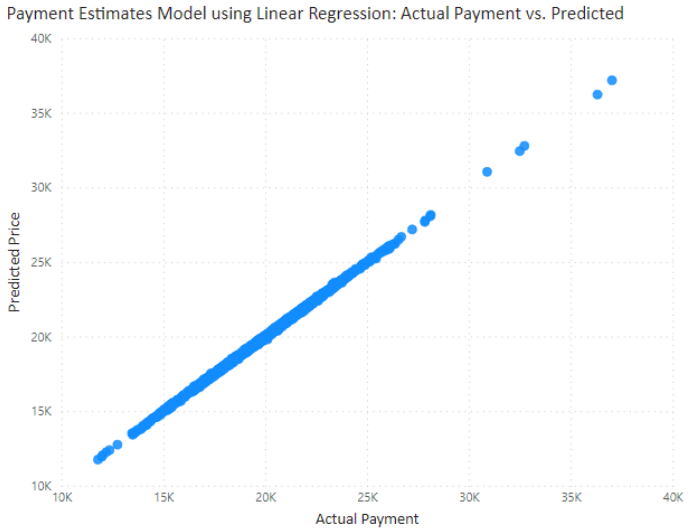
Payment and Value of Care

Do price and value of care always align?

Our remaining exploratory questions are directed towards better understanding the relationship and influence of value of care measures to price. Running a correlation matrix on our data, we can see moderately positive correlations between average payment and value of care measures including average efficiency ratio and average survival rate from the clinical outcomes data source as well as a weaker negative correlation between average payment and overall engagement score rating. While these correlations do not suggest strong relationships, we can see that value of care measures can impact price slightly or moderately.

	Total Population	*Publicly Insured Population*	***Average Payment***
Total Population	1		
Publicly Insured Population	0.983	1	
Average Payment	0.477	0.444	1
Overall Engagement Score Rating	-0.177	-0.195	**-0.314**
Count of Reported Hospitalizations	0.955	0.916	0.542
Average Complication Rate of Total Hip/Knee Arthroplasty	-0.058	-0.055	**-0.131**
Average Survival Rate of AMI, HF, and PN	0.262	0.255	**0.422**
Average Safety Ratio	0.100	0.100	**0.145**
Average Efficiency Ratio	0.186	0.174	**0.560**

As previously mentioned, our primary goal with our research project is to build a machine learning model that can best predict price with input of the associated features from the HVBP domain values. Our initial Payment Estimates model used linear regression to predict price using higher and lower estimates provided by the payment and value of care dataset and found a R^2 score of .999.



Ultimately, after further research, we determined that these higher and lower estimates were confidence interval estimates and therefore could not be used in future prediction models. The following machine learning models will be utilizing all previously mentioned features within the HVBP domains: efficiency, person and community engagement, safety, and clinical outcomes, and will not be trained or tested on higher estimate or lower estimate features.

Machine Learning: Regression Models (Lasso, Random Forest, & XGBoost)

Lasso Regression Payment Model

```
from sklearn.linear_model import Lasso

la_params = {'alpha': [0.0001,0.001,0.01,0.1,1,10,100]}

larg = Lasso()
la_grid = GridSearchCV(larg, la_params, scoring='r2', cv=5)
la_grid.fit(X_train, y_train)
y_preds = la_grid.predict(X_test)
print(r2_score(y_test, y_preds))
0.6420533626830127
```

Random Forest Regressor Model

```
from sklearn.ensemble import RandomForestRegressor

rf_params = { "n_estimators"      : [10,20,30],
              "max_features"      : ['sqrt', 'log2'],
              "min_samples_split" : [2,4,8],
              "bootstrap": [True, False],
            }

rfrg = RandomForestRegressor()
rf_grid = GridSearchCV(rfrg, rf_params, scoring='r2', cv=5)
rf_grid.fit(X_train, y_train)
y_preds = rf_grid.predict(X_test)
print(r2_score(y_test, y_preds))
0.6760349109986256
```

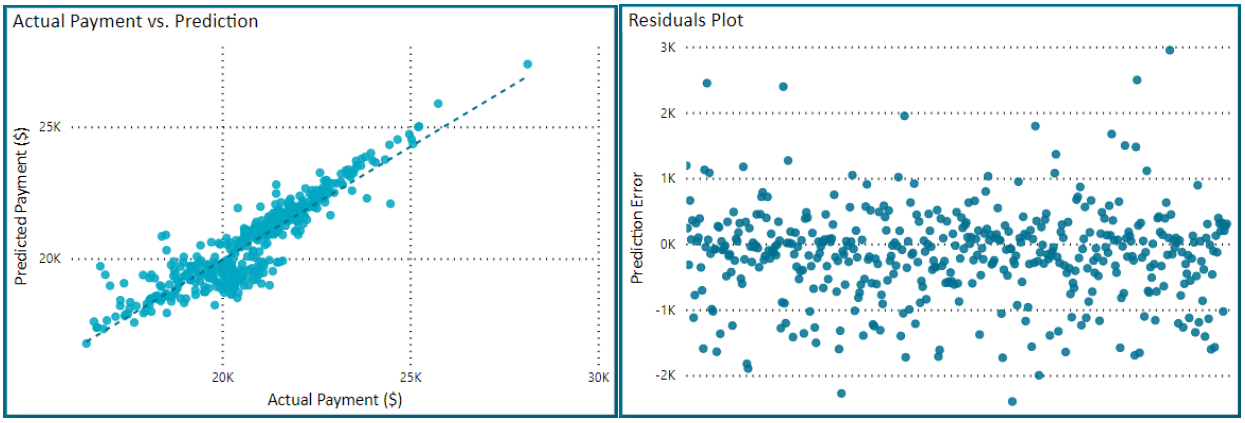
XGBoost Model

```
from xgboost import XGBRegressor

xg_params = {'nthread':[4],
             'objective':['reg:squarederror'],
             'learning_rate': [.03, 0.05, .07],
             'max_depth': [5, 6, 7],
             'min_child_weight': [4],
             'subsample': [0.7],
             'colsample_bytree': [0.7],
             'n_estimators': [500]
            }

xgrg = XGBRegressor()
xg_grid = GridSearchCV(xgrg, xg_params, scoring='r2', cv=2, n_jobs=5)
xg_grid.fit(X_train, y_train)
y_preds = xg_grid.predict(X_test)
print(r2_score(y_test, y_preds))
0.7830035159585188
```

XGBOOST MODEL:



As shown above, our XGBoost model had the highest R^2 score by .10 compared to our other regression models. This model accounts for about 80% of the variance of the dependent variable which implies there is correlation to be found in this data between payment and HVBP features. When evaluating the mean squared error (RMSE) of the model, we found that the model had an average residual of about $765, which we conclude is an acceptable error considering the average payment is about $20,322 per hospitalization. While considered a "black box" machine learning model, XGBoost is built to boost the forest of decision trees by combining smaller, weaker trees into larger, stronger ones and is a great fit for our vast features. The ability to tune hyperparameters by setting the maximum depth of each tree, the minimum child weight of a node, the gamma of the model and the number of columns used in each tree as well as the built-in cross validation on the learning rate of the model led us to determine that the XGBoost model was the most reliable choice in its ability to predict price.

Summary, Discussion, and Recommendations

While completing our detailed research and machine learning project, we have found that while correlations between all value of care features vary, our XGBoost model can reliably predict price. Our data analysis into the Census Bureau data on public insurance in the United States found a strong correlation between states' total population and publicly insured population. Four of the largest states were outliers when investigating correlation between total population and public insurance only versus VA public insurance and our analysis led to recommendations for possible increases in programming or advertising for other public insurance options for non-veteran residents as well as future research into specific correlations between VA-insured residents, payment estimates, and clinical outcomes specifically in the states of Texas and Florida. In terms of cost reduction, we found that higher efficiency scores did not lead to higher infection rates in hospitals concluding that there is not a visible risk to safety by hospitals trying to match the national median MSPB. However, because of the pandemic, many measures were compressed or not considered when determining HVBP measures so it would be critical to re-assess in future reporting periods to have a more accurate determination of impact between safety scores and efficiency scores. Overall, we have determined that payment and value of care does not always align but there is enough of a relationship among the value of care measures to reliably predict price when utilizing our XGBoost model compared to other regression models.

References

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Centers for Medicare and Medicaid Services (CMS). (2022). *Hospital Downloadable Database Data Dictionary: Centers for Medicare and Medicaid Services*. Retrieved from https://data.cms.gov/provider-data/sites/default/files/data_dictionaries/hospital/HospitalCompare-DataDictionary.pdf

Centers for Medicare and Medicaid Services (CMS). (2020). *Hospital Value-Based Purchasing (VBP) Program: How to Read Your Fiscal Year (FY) 2022 Baseline Measures Report*. Retrieved from https://www.qualityreportingcenter.com/globalassets/2020/02/iqr/hvbp-fy2022-baseline-report-help-guideupdated-002_vfinal508.pdf

Centers for Medicare and Medicaid Services (CMS). *About the Hospital VBP Program: Overview*. QualityNet. https://qualitynet.cms.gov/inpatient/hvbp

Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS). (2022, January 26). *Summary Analyses*. Retrieved from https://hcahpsonline.org/en/summary-analyses/

Hospital Value-Based Purchasing (HVBP) - Efficiency Scores. (2022, April 27). [Dataset]. https://data.cms.gov/provider-data/dataset/su9h-3pvj

Hospital Value-Based Purchasing (HVBP) - Person and Community Engagement Domain Scores (HCAHPS). (2022, April 27). [Dataset]. https://data.cms.gov/provider-data/dataset/avtz-f2ge

Hospital Value-Based Purchasing (HVBP) - Safety. (2022, April 27). [Dataset]. https://data.cms.gov/provider-data/dataset/dgmq-aat3

Hospital Value-Based Purchasing (HVBP) - Clinical Outcomes Domain Scores. (2022, April 27). [Dataset]. https://data.cms.gov/provider-data/dataset/pudb-wetr

How CMS Calculates the Patient Experience of Care (HCAHPS) Domain Score in the Hospital Value-Based Purchasing Program. (2018, February 7). Retrieved from https://hcahpsonline.org/globalassets/hcahps/vbp/hospital-vbp-domain-score-calculation-step-by-step-guide-february-2018.pdf

Joblib: Running Python Functions as Pipeline Jobs. (2020). Joblib Development Team. Retrieved from https://joblib.readthedocs.io/en/latest/generated/joblib.dump.html

Linking quality to payment. Centers for Medicare and Medicaid Services (CMS). Retrieved from https://data.cms.gov/provider-data/topics/hospitals/linking-quality-to-payment

Payment and value of care - Hospital. (2022, January 26). [Dataset]. https://data.cms.gov/provider-data/dataset/c7us-v4mf

Scikit-learn: Machine Learning in Python. (2011). Pedregosa et al., JMLR 12, pp. 2825-2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

U.S. Census Bureau, *2015-2019 5-Year American Community* Survey, Table S2704: Public Health Insurance Coverage by Type and Selected Characteristics (All States within United States). (2020) [Data set]. https://data.census.gov/cedsci/table?t=Health%20Insurance&g=0100000US%240400000&tid=ACSST5Y2020.S2704&tp=true