BAX 452 - MACHINE LEARNING

FINAL PROJECT

Cindy Jeon

Mia Lai

Sadhanha Anand

EXECUTIVE SUMMARY

In this project, we aimed to tackle the business problem of identifying patterns and actionable insights within a healthcare dataset focusing on obese and hypertensive patients aged between 40 and 75. Key findings reveal significant associations between patient demographics, medical prescription patterns (notably for Ozempic), and diagnostic categories. And we conduct a demographic-specific cost analysis to determine which patient groups are generating the highest healthcare costs and track the cost trends for medications like Ozempic over time. Based on the analysis, we recommend targeted healthcare interventions and patient education programs, particularly focusing on the underrepresented groups identified in the Ozempic prescription data. These actions are expected to improve patient outcomes and optimize healthcare resource allocation.

INTRODUCTION

Business Problem: The healthcare industry faces challenges in effectively targeting and treating obese and hypertensive patients. Misallocation of resources and lack of targeted care result in suboptimal patient outcomes and increased healthcare costs. Addressing these issues requires deep data-driven insights into patient demographics and treatment patterns.

Objectives: The primary goal is to analyze patient data to uncover trends, assess the effectiveness of medications like Ozempic, and identify areas for improvement in patient care and resource allocation.

DATASET DESCRIPTION

Source: The dataset originates from medical records and prescription data of obese and hypertensive patients aged 40 to 75.

Variables:

- ➢ Patient demographic information (age, gender, etc.)
- ➢ Medical diagnosis codes
- ➢ Medication prescription records (including Ozempic NDC numbers)
- ➢ Visit records and healthcare provider details

Preprocessing Steps: Included cleaning missing values, encoding categorical variables, and extracting relevant features like diagnosis codes.

METHODOLOGY

Exploratory Data Analysis (EDA): We analyzed demographic distributions, diagnosis frequencies, and medication prescription patterns. Visualizations such as bar charts and scatter plots were employed to uncover trends and outliers. Based on the plots, we were able to come up with the following insights:

Patient Demographics and Locations

- ➢ The majority of patients come from a narrow range of zip codes, suggesting geographic hotspots for the studied conditions.
- ➢ Patients predominantly fall within the 60-70 age range, highlighting this as a critical period for the onset or management of the conditions studied.
- ➢ California (CA) has a significantly higher number of patients compared to other states, indicating a possible regional prevalence of the health conditions under study.

Healthcare Provider Insights

A few healthcare organization identifiers (hco_npi) and revenue center codes show unusually high frequencies, suggesting specific healthcare providers or services are more involved in the patients' care.

Financial Aspects

Line charges, claim charges, and amounts allowed are mostly concentrated near lower values, implying that the majority of the provided healthcare services are of lower cost.

Diagnosis and Treatment Patterns

- ➢ The most common diagnoses include diabetes (E11) and hypertension (I10), followed by hyperlipidemia (E78), encounter for immunization (Z00), and hypothyroidism (E66), indicating prevalent chronic conditions within the patient population.
- ➢ A higher prevalence of obesity diagnoses is observed among females (58.72%) compared to males (41.28%), suggesting gender disparities in either the reporting or incidence of obesity.
- ➢ The age distribution for obesity diagnosis peaks in the 50-60 age group, with a noticeable increase in diagnoses starting from age 40, which remains high up to the 70s before declining.

Gender Disparities in Hypertension Diagnosis

The line chart illustrating the number of patients diagnosed with hypertension by gender over months shows that females consistently have higher counts than males, indicating a potential gender disparity in hypertension prevalence or reporting.

Medication Prescription Patterns

The bar chart comparing patients who received Ozempic versus those who did not shows that a vast majority fall into the 'Others' category. This suggests that Ozempic is prescribed to a relatively small portion of patients, indicating either underutilization or selective prescription practices for this medication.

Model Selection and Evaluation: We explored various statistical and machine learning models to predict patient outcomes based on their medical history and demographics. Models included LassoCV, LogisticRegressionCV, Random Forest and Neural Networks, chosen for their ability to handle high-dimensional data. Performance was assessed using cross-validation, accuracy score, mean squared error, and R2 score, to ensure robustness and reliability of the findings.

Findings: Our analysis indicated significant variability in medication prescriptions across different patient groups. For instance, Ozempic was prescribed significantly less frequently than other medications, suggesting potential underutilization.

Interpretation: These results suggest potential gaps in treatment approaches, especially concerning the use of Ozempic, and highlight the necessity for targeted intervention strategies.

First Model selection : Double lasso

- Model evaluation and interpretation:

  1) R^2: The value is 14.82%, CV average value is 10.24%, suggesting that the model explains 14.582% of the variability in the binary outcome, which is not very high. This implies that other unaccounted-for factors are influencing whether a patient is diagnosed with obesity or hypertension on the first diagnosis.

  2) MSE: The value is 0.1693, CV average value is 0.177. The MSE here suggests that there is some level of error in the model's predictions.

  These two criterias suggest that the treatment effect of Ozempic is not significant.

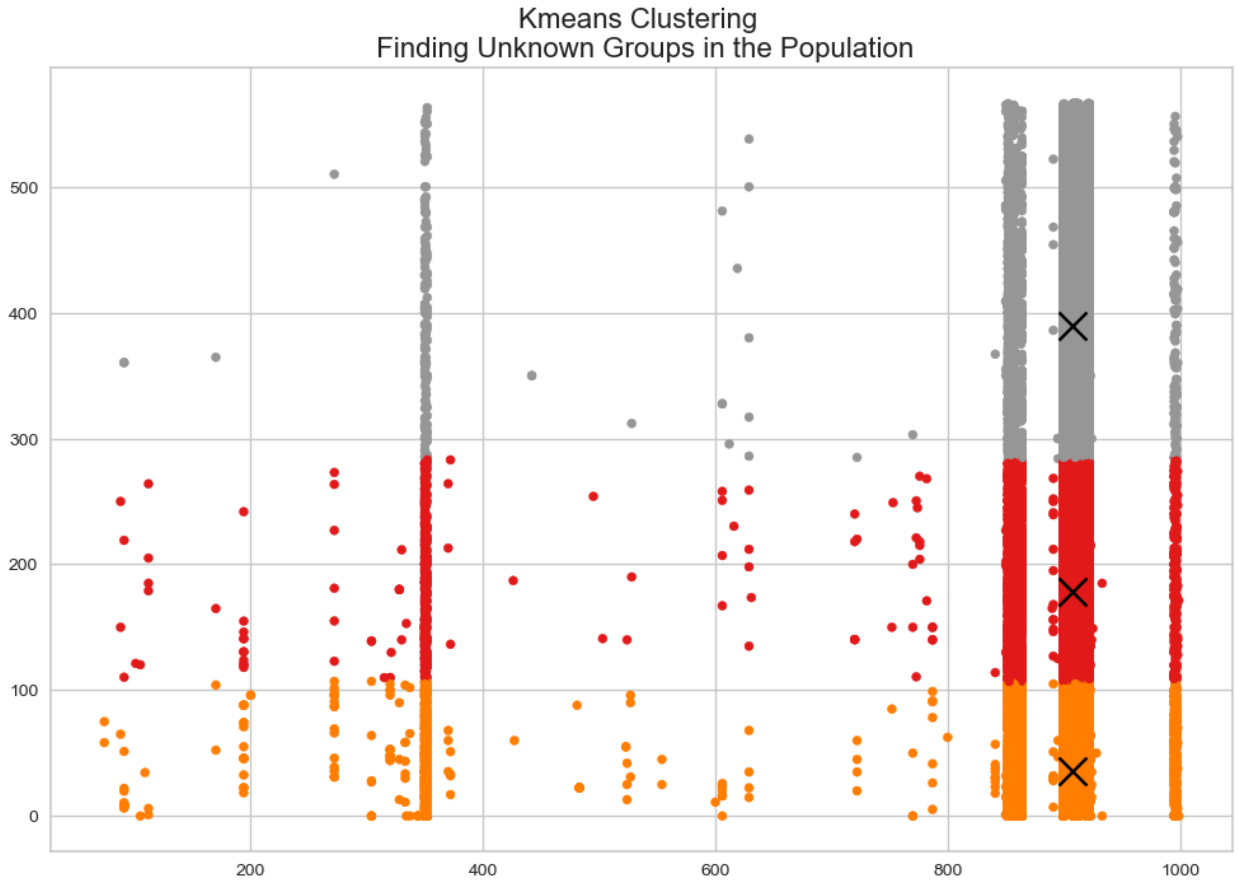
Second Model selection : Random Forest and Classification

- Model evaluation and interpretation: we used both random forest and classification to measure how accurately the model can classify cured people based on variables, such

as 'patient_state', 'patient_age', 'patient_gender', 'diag_list', 'diag_2', 'diag_3', 'diag_4', 'diag_5', 'ndc = 1

MSE is 0.080 indicates this model is very well performing with random forest. Accuracy is 91% and MSE is 0.087 for the classification and this model is very well performing

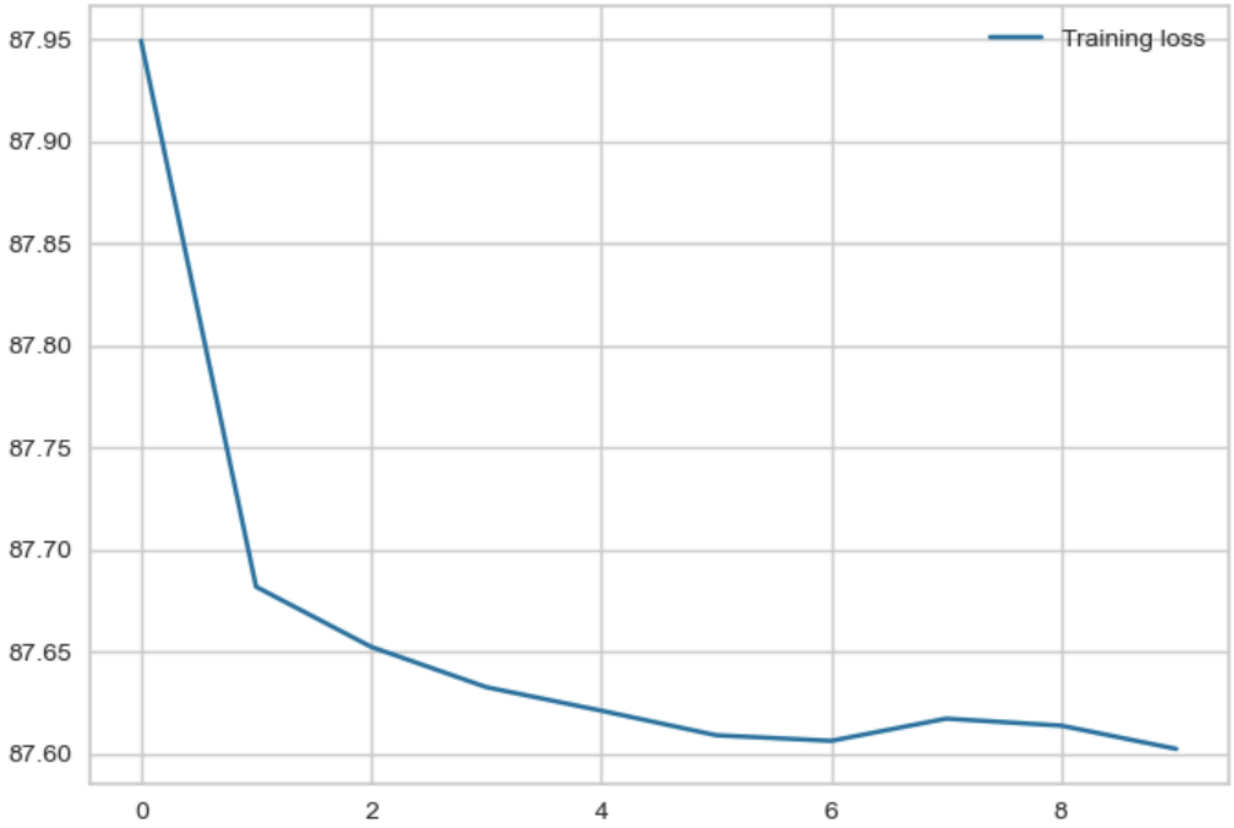
Third Model selection : K-means Clustering and Neural Network (Tensor Flow)



- Model evaluation and interpretation:

  1) K-means Clustering: The graph depicts several clusters identified by the K-means algorithm, each represented by a different color. The black "X" marks presumably represent the centroids of these clusters. There may be significant differences in claim charges based on patients' locations, as suggested by the clusters' alignment along the 'short zip' axis. The spread of 'claim charge' values within each cluster varies, but there

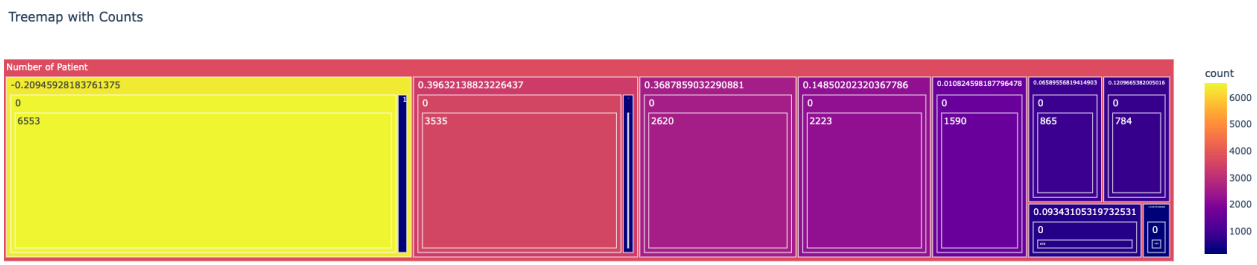
appears to be a cluster (in red) with particularly high 'claim charges'. This could represent a group of patients with more expensive medical costs.



2) TensorFlow: Initially, there is a steep decrease in the training loss, which suggests that the model is quickly learning from the training data. The plot shows low variability in loss after the initial epochs, which could be a good sign of model stability.

ADDITIONAL INSIGHTS

1) Diagnose group:



The largest area is colored yellow, representing the diagnostic code 900.0, with the highest patient count of 6553. This indicates Los Angeles has the highest number of patients. Higher patient counts in larger cities like Los Angeles could be reflective of both higher population density and a broader availability of healthcare services leading to more diagnoses being recorded. Comparatively high patient counts in smaller cities, like Indio, may signal specific health challenges or higher prevalence of certain conditions that warrant further investigation.

2) **Logistic Regression (logistic regression to predict Ozempic treatment effect based on the demographic and diagnosis information)**

1. Precision: the value is 92.99%, which means when the model predicts that a patient is cured, it is correct about 92.99% of the time. This high precision suggests that the presence of the treatment is a strong predictor for a cured outcome in the predicted instances.

2. Recall (F1 Score): the value is 91.66%, suggesting the model is able to correctly identify 91.66% of the patients who are actually cured. High recall in the context of treatment effect suggests that Ozempic is effective for a large proportion of the patients.

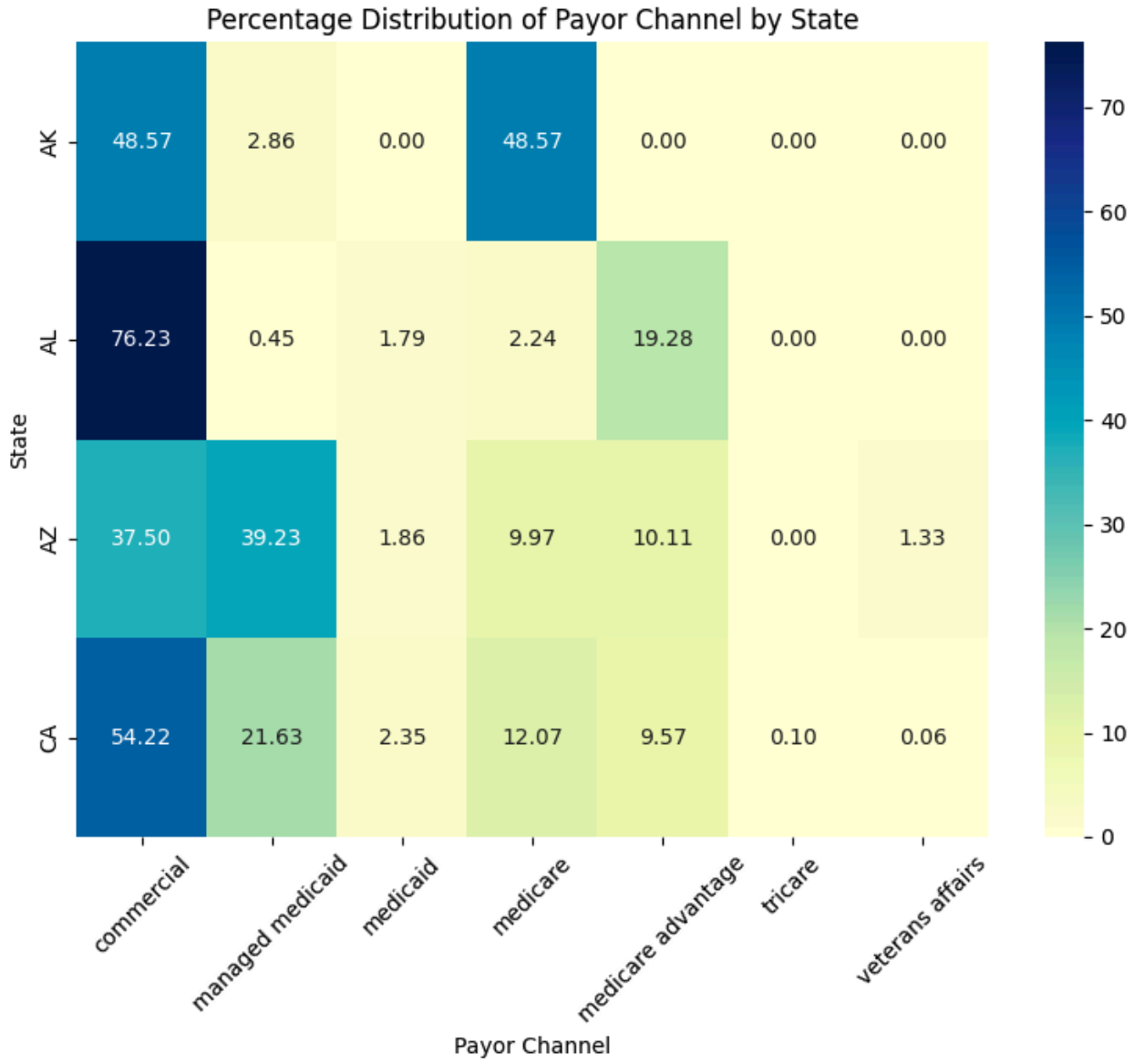
3. AUC-ROC: the value is 86.09%, which means the model has a good ability to distinguish between patients who are cured and those who are not.

The logistic regression model shows strong performance across precision, recall, and AUC-ROC metrics. However, there is room for improvement, as indicated by the number of false positives and false negatives.

- Comparison:

Based on these metrics, the logistic regression model appears to be significantly more effective at predicting the outcome variable in this dataset compared to the double lasso regression model. However, the goal is the inference about the importance of specific variables (treatment effect); double lasso regression may still provide valuable insights despite its lower discriminative performance.

3) Percentage Distribution of Payor Channel by State



- When filtering the data for only individuals aged 40 to 75 with obesity and hypertension, although California (CA) has the highest number of users of Ozempic, the proportion of those using commercial payers is 54%. In contrast, Alabama (AL) has a 76% commercial payor proportion, indicating that a higher percentage of individuals in AL may pay for Ozempic without

insurance coverage. Based on this, we can't say that we are confident enough that usage of ozempic is correlated with place/state/income/education level.

RECOMMENDATIONS

Based on our findings, we recommend:

- ➢ Increased awareness and education on Ozempic for eligible patients.
- ➢ Further investigation into the barriers to effective medication prescription.
- ➢ Development of personalized treatment plans based on patient demographics and medical history.
- ➢ Use the model's predictions to prioritize preventive health measures in high-cost clusters to mitigate future expenses.
- ➢ Develop dynamic pricing models for treatments like Ozempic, adjusting for factors revealed in the clustering such as geographic and demographic variations.

CONCLUSION

This project shed light on crucial aspects of healthcare provision for obese and hypertensive patients. Through rigorous data analysis, we identified key areas for improvement and actionable insights that can lead to enhanced patient care and optimized healthcare strategies.

APPENDIX

1. Code: https://drive.google.com/file/d/1HqoCM6WCU_TtKyeQ_MJTXf7R9foaxccz/view?usp=sharing