INSURANCE HOLDERS OUT OF POCKET PRICE PREDICTIONS USING MACHINE LEARNING ALGORITHMS

Anshul Chimnani - 10609598

Supervisor: Prof. Paul Laird

Dublin Business School

This dissertation is submitted for the degree of
Master of Science in Data Analytics

January 2024

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Anshul Chimnani - 10609598
January 2024

Acknowledgements

I would like to acknowledge my supervisor Mr.Paul Laird for his constant support and motivation during the course of the project. I appreciate his acceptance of all of my ideas and his ability to keep me focused.

I express my gratitude to all my family and friends for their unwavering support throughout this period.

Abstract

This research project explores the field of healthcare economics, utilizing machine learning methods to forecast the out-of-pocket costs for healthcare services. Using a dataset covering the years 2017 to 2021, the project seeks to create a strong prediction model that can accurately forecast out-of-pocket spending for the year 2022 according to the diagnosis of the patient. The dataset is carefully selected and organized, including a wide range of characteristics such as patient demographics, healthcare price details, age band, and Diagnosis of the patient, etc.

The project adheres to a methodology that begins with the gathering and preparation of data. The dataset is divided into three separate sets - a training set consisting of data from 2017 to 2019, a test set representing the year 2020, and a test validation set covering the year 2021. The model's capacity to identify patterns and dependencies within the data is improved by meticulous feature selection and engineering.

An assortment of machine learning algorithms is assessed, encompassing traditional linear regression models as well as more intricate ensemble techniques such as random forests and support vector regression.

Performance evaluation is an essential aspect of research, which entails thorough assessments on both the validation and test sets. Metrics such as mean absolute error and mean squared error are used to measure the accuracy and dependability of the model. Afterwards, the model that has been trained is used on the 2022 dataset to produce forecasts for out-of-pocket prices.

This research enhances the comprehension of the determinants impacting healthcare costs and offers a prognostic instrument for stakeholders in the healthcare sector. The consequences have a broader reach beyond the academic realm, providing practical insights that are valuable for policy-makers, insurers, and healthcare providers. Furthermore, the paper examines the ethical implications of using predictive models in healthcare, highlighting the significance of being open and responsible.

Furthermore, the study utilizes neural networks to enhance forecasting abilities by evaluating 2021. The machine learning pipeline incorporates a deep learning architecture, enabling the model to capture complex non-linear correlations in the data. The neural network archi-

tecture is extensively trained on the training dataset to adapt to the intrinsic complexity of healthcare price dynamics. The combination of conventional machine learning and neural network approaches strengthens the project's dedication to utilizing state-of-the-art methods for accurate and detailed projections of healthcare expenses that individuals have to pay themselves.

Ultimately, this research project combines machine learning techniques and neural networks with healthcare economics to create a predictive model that can improve decision-making in the healthcare industry. The research and ideas offered here contribute to the continuing discussion about the affordability and accessibility of healthcare services.

Keywords: Out-of-pocket prices, Healthcare economics, Machine learning, Predictive modeling, Dataset, Feature engineering, Validation, Neural networks, Deep learning, Healthcare expenses.

Table of contents

List of figures

Chapter 1

Introduction

1.1 Background and Motivation

Out-of-pocket costs have a tremendous impact on the financial picture of modern healthcare systems, placing a substantial strain on people and families. A comprehensive take on predictive modeling for out-of-pocket charges is required due to the rising expenses of healthcare services and the complex interaction of multiple elements. The motivation behind this research endeavor stems from the urgent requirement for sophisticated methods to precisely predict these expenditures.

The escalating expenses of healthcare have extensive consequences, affecting the accessibility and affordability for various demographic cohorts. The limitations of conventional pricing methods highlight the necessity of integrating sophisticated approaches, such as machine learning and neural networks, to identify intricate patterns within healthcare data. The research is motivated by the convergence of healthcare economics and predictive modeling, with the aim of connecting theoretical findings to practical applications.

The driving force behind this research is the possibility of completely transforming decision-making procedures in the healthcare sector. Precise forecasts of patient's expenses enable policymakers, insurers, and healthcare providers to develop tactics that improve clarity and reduce financial hardships for individuals. This project aims to provide significant insights into the complexities of healthcare expenditure, with the goal of contributing to a wider discussion on the accessibility and affordability of healthcare.

Moreover, the incorporation of neural networks enhances the complexity of the prediction model, enabling a more profound comprehension of non-linear correlations within the data. The motive is not solely limited to academic pursuits; it encompasses the tangible effects of empowering stakeholders to make well-informed decisions, efficiently allocate resources, and eventually cultivate a fairer healthcare ecosystem.

1.2 Research Objectives

Objectives of the research:

The objective of this research project is to fill important knowledge gaps on the costs that individuals pay directly for healthcare services, by utilizing advanced predictive modeling approaches. The main goals can be defined as follows:

Create a Predictive Model: The primary objective is to build a resilient predictive model that can precisely anticipate out-of-pocket costs for healthcare services and be more efficient for the diagnosis of the patient. The study aims to identify intricate linkages within the data that conventional models may fail to detect by employing machine learning methods and neural networks.

Employ Historical Data: To improve the effectiveness of the model, we will utilize historical data from the years 2017 to 2021. The research will methodically examine this dataset, integrating diverse characteristics such as patient demographics, healthcare provider information, and economic indicators, to identify patterns and trends that impact out-of-pocket spending.

Assess and verify the model: The project prioritizes thorough assessment and verification procedures. The model's accuracy and generalization will be evaluated using performance metrics such as mean absolute error and mean squared error. This evaluation will be conducted on both the validation set (2021) and the test set (2020).

Integrate Neural Networks: Specifically, the study aims to enhance conventional machine learning approaches by including neural networks. The project seeks to uncover complex non-linear relationships within healthcare data by investigating deep learning architectures. This will enhance the predictive model by providing a more detailed and precise understanding.

Enhance the process of making informed decisions in the healthcare sector: In addition to its academic nature, this research has a practical focus, aiming to provide valuable insights for decision-making in the healthcare industry. Precise forecasts of the costs that individuals have to pay directly have the capacity to assist policy-makers, insurers, and healthcare providers in formulating approaches to improve the affordability and availability of healthcare.

The project aims to provide useful insights to the subject of healthcare economics. The project seeks to enhance comprehension of healthcare expenditure dynamics and their wider consequences by offering an advanced forecasting tool and a thorough examination of factors that impact out-of-pocket expenses.

These aims jointly aim to advance predictive modeling in healthcare and provide practical answers to current difficulties in the healthcare economic landscape, with the goal of making a significant effect.

1.3 Research Questions

Listed below are potential research concerns that we may address later in the project:

1. How effectively can machine learning algorithms predict out-of-pocket prices for healthcare services using historical data from 2017 to 2021?

2. What features and variables have the most significant impact on predicting out-of-pocket expenses in healthcare, and how can they be leveraged to enhance model accuracy?

3. How well does the predictive model generalize to new data, as demonstrated by its performance on the validation set (2020) and the test set (2021)?

4. What are the key insights derived from the analysis of model predictions, and how do they contribute to our understanding of the factors influencing out-of-pocket prices in healthcare?

5. In what ways can the developed predictive model inform decision-making processes for policy-makers, insurers, and healthcare providers to enhance the affordability and accessibility of healthcare services?

6. What ethical considerations, including privacy, data security, and bias, should be addressed in the deployment of predictive models for healthcare economics, and how can these concerns be mitigated?

. To what extent does the predictive model align with real-world out-of-pocket expenses in the year 2022, and how reliable are the projections for future healthcare expenditures?

9. What limitations exist in the current predictive model, and what areas require further refinement or exploration for future research in the domain of healthcare economics and predictive modeling?

10. How can the findings of this research project contribute to the broader discourse on healthcare accessibility and affordability, and what practical implications do they hold for stakeholders in the healthcare industry?

These research questions guide the investigation, allowing for a comprehensive exploration of predictive modeling in healthcare economics while addressing key aspects of accuracy, generalization, ethical considerations, and practical applications.

Chapter 2

Literature Review

2.1 Healthcare Economics

The subject of healthcare economics is constantly changing, and predictive modeling has become increasingly important. It offers useful insights into the dynamics of costs, allocation of resources, and the design of policies. This literature review consolidates significant contributions in the field of predictive modeling for healthcare expenses that individuals pay directly.

2.2 Out-of-Pocket Spending in Healthcare

Previous studies have found various factors that impact the amount of money individuals spend out of their own pockets. Finkelstein et al. (2020) emphasize the influence of insurance coverage and individual attributes on healthcare expenses. It is crucial to include these aspects in predictive models in order to get precise estimates.

2.3 Predictive Modeling in Healthcare

The utilization of predictive modeling in the healthcare sector has shown substantial expansion in recent years Obermeyer and Emanuel (2016) and Chen et al. (2016) have conducted studies that show the effectiveness of machine learning algorithms in accurately forecasting healthcare outcomes and expenses. These findings provide a fundamental framework for investigating prediction models in relation to out-of-pocket spending.

2.4 Machine Learning in Healthcare Economics

Diverse machine learning methods have been utilized in the field of healthcare economics. Song et al. (2021) demonstrate the use of decision trees and random forests to forecast healthcare expenses, highlighting their capacity to be easily understood and their high level of accuracy. The present project's selection process is influenced by these automated selections. Incorporating neural networks into healthcare prediction modeling has demonstrated potential in capturing complex patterns LeCun et al. (2015) and Rajkomar et al. (2019) explore the use of deep learning architectures in healthcare data, emphasizing their ability to effectively handle intricate and non-linear interactions.

2.5 Generalization and Validation

Ensuring the generalization of prediction models to new data is a crucial feature. Harutyunyan et al. (2019) and Caruana et al. (2015) emphasize the significance of rigorous validation procedures and examine methods for assessing the performance of models on unfamiliar data. These coincide with the validation approach used in this investigation.

2.6 Actual-World Applications and Impact

Chen et al. (2020) delve into the convergence of predictive modeling and its tangible effects on the actual world. These studies emphasize the practical consequences of using predictive models in healthcare decision-making, aligning with the project's goal of providing information to stakeholders.

2.7 Ethical Implications of Prediction Algorithms

As predictive models become more popular, ethical considerations become extremely important. O'neil (2017) and Mittelstadt et al. (2016) examine the ethical ramifications of algorithmic decision-making in healthcare, with a particular focus on concerns such as bias, transparency, and responsibility.

References

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730.

Chen, I. Y., Joshi, S., and Ghassemi, M. (2020). Treating health disparities with artificial intelligence. *Nature medicine*, 26(1):16–17.

Chen, J., Vargas-Bustamante, A., Mortensen, K., and Ortega, A. N. (2016). Racial and ethnic disparities in health care access and utilization under the affordable care act. *Medical care*, 54(2):140.

Finkelstein, A., Zhou, A., Taubman, S., and Doyle, J. (2020). Health care hotspotting—a randomized, controlled trial. *New England Journal of Medicine*, 382(2):152–162.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216.

O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Song, J., Gao, Y., Yin, P., Li, Y., Li, Y., Zhang, J., Su, Q., Fu, X., and Pi, H. (2021). The random forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms. *Risk Management and Healthcare Policy*, pages 1175–1187.

Chapter 3

Data Collection and Cleaning

3.1 Data Source

In this project, I used the sample dataset available on the website of Health Care Cost Institute (HCCI). The dataset provides a detailed range of years, agebands, diagnoses, patient types, procedure types, out-of-pocket price, medical costs, and healthcare usage (utilization). The main sources of HCCI comprise of official health department reports and surveys administered by various healthcare institutions.

The dataset is intricately organized, encompassing factors such as patient demographics, utilization, and price per person. Data cleaning procedures encompass the management of missing values, the standardization of units, and the assurance of privacy compliance for patient data. The cleaned dataset obtained serves as a solid basis for performing exploratory data analysis, comprehending healthcare spending trends, and extracting insights to enhance the affordability and accessibility of healthcare.

3.2 Data Description

Descriptions of the columns

3.2.1 yr:

This column provides the year in which the claims occurred. It is particularly significant for dividing the information into two subsets: one for the year 2021 and another for the years 2017 to 2020.

3.2.2 Age band (age_band):

The term "age band" pertains to the specific age range of a patient when they receive healthcare services. It is used to gather data on the demographic distribution of patients. (AGE between 0 and 17 = 1, AGE between 18 and 24 = 2, AGE between 25 and 34 = 3, AGE between 35 and 44 = 4, AGE between 45 and 54 =5, AGE between 55 and 64 =6, if age >= 65 = 7)

3.2.3 Gender (sex)

: Specifies the biological or self-identified gender of the individual.

3.2.4 Patient type (hcci_detcat_1):

Indicates the specific category of the patient such as Inpatient, Outpatient, Professional Services, or Prescription Drugs.

3.2.5 Patient Diagnosis (hcci_detcat_2):

This column shows the diagnosis of the patient in different categories such as Eye, Circulatory, Newborn, Male Reproductive, Endocrine, etc.

3.2.6 Procedure (hcci_detcat_3):

This column refers to the precise medical procedure or service administered to the patient, encoded for the purpose of standardization.

3.2.7 Out-of-Pocket Spending per Person: (oop_per_person

The calculation involves aggregating the monetary contributions made by individuals (members) for healthcare services, which include copayments, co-insurance, and deductibles. The computation is performed for submitted claims and is then divided by the mean count of individuals covered by employer-sponsored private health insurance (ESI).

3.2.8 Price (Mean Price per Service)

: The prices are determined by dividing the total expenditure by the total utilization for each service category, resulting in the average price per service. It denotes the mean sum paid by both the payer and the patient directly for a specific healthcare service.

3.2.9 Per capita utilization (util_per_100

: Utilization rates are determined by aggregating the number of admissions, procedures, and filled prescription days per 1,000 covered individuals. A utilization measure is calculated for each service category, yielding a count of utilization per person for each category. This figure is then multiplied by 1,000.

3.2.10 Out of pocket price (oop_price) :

Out-of-pocket prices are determined by each individual based on diagnosis.

3.3 Importing Libraries

3.3.1 Scikit-learn

I executed the command "pip install scikit-learn" to install the scikit-learn library. Scikit-learn is an extensive Python library for machine learning. The platform offers user-friendly and effective tools for analyzing and modeling data, encompassing a wide range of methods for regression, classification, clustering, and preprocessing approaches. In my code, I have employed the following techniques: linear regression (LinearRegression), random forest regression (RandomForestRegressor), support vector regression (SVR), data splitting (train-test-split), and evaluation metrics (mean squared error and $r2$-score).

3.3.2 Category-Encoders

I installed the category-encoders package with the command "!pip install category-encoders". Category Encoders is a software package designed to encode category variables in machine learning models. In my code, I am utilizing the OneHotEncoder class from both the scikit-learn and category-encoders libraries. By utilizing this function, I convert categorical characteristics in my dataset into a format suitable for machine learning models, which is an essential process in data preparation.

3.3.3 Pandas

Used the pandas library and import it as pd. Pandas is a robust library for manipulating and analyzing data. The library offers data structures such as DataFrame to facilitate effective data management and manipulation. In my code I will utilize the pandas library to load and manipulate the dataset.

3.3.4 Scipy.stats

Used the "import" command to bring the "f-oneway" function from the "scipy.stats" module. SciPy is a software library designed for the purpose of doing scientific and technical computations. The f-oneway function in the scipy.stats module is utilized for doing one-way analysis of variance (ANOVA), a statistical technique used to compare means across various groups.

3.3.5 Seaborn and Matplotlib

Use the "import" command to bring in the seaborn and matplotlib libraries. Import the pyplot module from the matplotlib library and alias it as plt. Seaborn and Matplotlib are libraries specifically designed for data visualization. Seaborn enhances the functionality of Matplotlib by offering a user-friendly interface for generating visually appealing statistical visualizations.

3.3.6 TensorFlow (Keras)

TensorFlow is an open-source library, while Keras is a neural networks API that operates at a higher level and is built on top of TensorFlow.I have included the Sequential class from the Keras library in the cose. This class is commonly utilized to construct neural network models in a step-by-step manner.

By utilizing these libraries, my code takes advantage of proven tools for manipulating data, conducting statistical analysis, developing machine learning algorithms, and creating neural networks. This simplifies the process of constructing and assessing prediction models.

3.4 Data Cleaning

The data cleaning process in my project has two essential steps. Initially, it applies a filtering process to extract and separate data specifically from the year 2021, to conduct a targeted study. Simultaneously, it eliminates entries that correspond to the year 2021 from the original dataset. This stage guarantees the exclusion of entries from that particular year in order to provide a more refined dataset for further research. Furthermore, the code detects and exhibits distinct years found in the altered dataset, offering valuable observations into the chronological arrangement of the remaining data. This procedure facilitates the improvement of the dataset by eliminating data that is exclusive to the year 2021, hence boosting its overall quality.

Chapter 4

Methodology

4.1 Hypothesis Testing

4.1.1 Overview

Hypothesis testing is a statistical technique employed to conclude about population parameters using data obtained from a sample. The procedure entails constructing a null hypothesis (H0) and an alternative hypothesis (H1), which represent significant differences in the population. The null hypothesis proposes no difference, whereas the alternative hypothesis proposes a substantial difference. Researchers gather and examine sample data to see if there is sufficient evidence to reject the null hypothesis in support of the alternative hypothesis. This is accomplished by computing a test statistic, like as a t-test or z-test, that measures the discrepancy between the observed data and the expected outcome according to the null hypothesis. The p-value is an essential element in the process of hypothesis testing. The term "p-value" denotes the probability of detecting a test statistic as extreme as the computed one, under the assumption that the null hypothesis is valid. A smaller p-value signifies more compelling evidence against the null hypothesis, leading researchers to reject it. The significance level, commonly represented as α , is the preset threshold used to assess whether to reject the null hypothesis. Typically, a significance level of 0.05 is employed. When the p-value is smaller than α , researchers discard the null hypothesis, indicating that the observed data offers sufficient evidence to substantiate the alternative hypothesis. If the p-value is bigger than α, it indicates that there is not enough evidence to reject the null hypothesis. Hypothesis testing enables researchers to make inferences about population parameters, offering a methodical and unbiased approach to evaluate the accuracy of assertions using data from a sample. It is extensively employed in diverse fields to make well-informed decisions and derive dependable insights from experimental or observational investigations.

4.1.2 Hypothesis test 1

Question - Is there a significant difference in out-of-pocket spending between different healthcare patient types?

This hypothesis test is examined using a statistical technique called Analysis of Variance (ANOVA). This is used to determine if there is a substantial difference in out-of-pocket spending among different healthcare patient categories. This statistical analysis puts a particular emphasis on out-of-pocket expenses for various healthcare patient categories. Let's discuss the code in detail:

```
# Group data by health care categories and calculate mean oop_per_person
healthcare_categories = data.groupby('hcci_detcat_1')['oop_per_person'].mean()

# Perform one-way ANOVA to test for significant differences in oop_per_person among categories
f_stat, p_value = f_oneway(*[group['oop_per_person'] for name, group in data.groupby('hcci_detcat_1')])
```

Fig. 4.1 Performing ANOVA test

The dataset is being aggregated using the column 'hcci_detcat_1', which represents different types of healthcare patients. Subsequently, this data is then grouped by the mean of the 'oop_per_person' column by using the function "groupby", which represents the average amount of money spent out of pocket for healthcare patients categories. The services addressed are Inpatient care, Outpatient care, Professional services, and Prescription drugs. The result is stored in the healthcare_categories variable, which is a Pandas Series containing the average amount of money spent out-of-pocket as values, with healthcare categories serving as the index.

Hypothesis Testing: The null hypothesis **(H0)** claims that there is no statistically significant difference in out-of-pocket spending across healthcare categories specifically Patient type in this case. The alternative hypothesis **(H1)** claims the existence of a significant difference.

The ANOVA F-statistic and p-value are calculated and shown in the output display. The F-statistic is the ratio of variance between group means to variance within groups. The p-value represents the probability of receiving these outcomes if the null hypothesis is supported.

```
print("Hypothesis 1: Is there a significant difference in out-of-pocket spending between different health care categories?")
print(f"ANOVA F-statistic: {f_stat}")
print(f"P-value: {p_value}")
if p_value < 0.05:
    print("Conclusion: There is a significant difference in out-of-pocket spending among health care categories.")
else:
    print("Conclusion: There is no significant difference in out-of-pocket spending among health care categories.")
```

```
Hypothesis 1: Is there a significant difference in out-of-pocket spending between different health care categories?
ANOVA F-statistic: 210.00664952958343
P-value: 2.2382258010103047e-134
Conclusion: There is a significant difference in out-of-pocket spending among health care categories.
```

Fig. 4.2 Results of ANOVA (Hypothesis 1)

Result: The code generates a conclusion based on the p-value. If the p-value is below 0.05 (a commonly used significance level), it indicates that the null hypothesis should be rejected in favor of the alternative hypothesis. This leads to the conclusion that there is a significant difference in out-of-pocket expenses across different healthcare categories.

In essence, the code serves as a statistical analysis tool that allows us to determine if there are significant differences in out-of-pocket spending across various healthcare categories, using the given data.

4.1.3 Hypothesis test 2

Hypothesis 2: Do surgical procedures have higher average prices compared to medical procedures? The hypothesis being tested in this part is an evaluation hypothesis, specifically a two-sample hypothesis. More precisely, it is a hypothesis that compares the means of two separate samples (one from surgical procedures and the other from medical procedures) in order to identify whether there exists a statistically significant difference in their average prices.

To further analyze the subject:

```
# Calculate mean prices for surgical and medical procedures
surgical_mean_price = data[data['hcci_detcat_3'] == 'S']['price'].mean()
medical_mean_price = data[data['hcci_detcat_3'] == 'M']['price'].mean()
```

Fig. 4.3 Hypothesis 2 test

The null hypothesis **(H0)** states that there is no statistically significant difference in the mean prices of surgical and medical treatments. Alternative Hypothesis **(H1)** states that Surgical operations have a greater mean cost in comparison to medicinal procedures.

```
print("Hypothesis 2: Do surgical procedures have higher average prices compared to medical procedures?")
print(f"Mean price of surgical procedures: {surgical_mean_price}")
print(f"Mean price of medical procedures: {medical_mean_price}")
if surgical_mean_price > medical_mean_price:
    print("Conclusion: Surgical procedures have a higher average price compared to medical procedures.")
else:
    print("Conclusion: There's no clear evidence that surgical procedures have a higher average price compared to medical procedures.")
```

```
Hypothesis 2: Do surgical procedures have higher average prices compared to medical procedures?
Mean price of surgical procedures: 49283.057800224466
Mean price of medical procedures: 22711.838377723972
Conclusion: Surgical procedures have a higher average price compared to medical procedures.
```

Fig. 4.4 Hypothesis 2 test output

Analysis and Final Remarks: Subsequently, the average costs of surgical and medical treatments are compared. If the average price of surgical procedures exceeds the average price of medical procedures, then the conclusion can be stated as "Surgical procedures have a higher average price compared to medical procedures." If there is insufficient data to support the notion that surgical procedures exhibit a higher mean cost, the logical conclusion is that "There is no definitive evidence to suggest that surgical procedures have a higher average price in comparison to medical procedures." The given output indicates that the average price of surgical procedures (49283.06) is certainly greater than the average price of medical procedures (22711.84), therefore suggesting that surgical treatments have a higher mean price in comparison to medical procedures.

4.1.4 Hypothesis test 3

Hypothesis 3: Is there a difference in utilization rates between genders in nervous system-related cases? The objective of Hypothesis 3 is to examine whether there is a significant difference in utilization rates between genders for the cases of the nervous system. The data is selectively refined to primarily include cases related to the neurological system. Subsequently, the utilization rates are calculated for both genders.

The null hypothesis **(H0)** states that there is no statistically significant difference in utilization rates between genders for cases related to the nervous system. Alternative Hypothesis **(H1)**: There exists a statistically significant difference in the utilization rates across genders for cases related to the neurological system.

```
[13] nervous_system_data = data[data['hcci_detcat_2_label'] == 'Nervous System']
     gender_utilization = nervous_system_data.groupby('sex')['util_per_1000'].mean()

     print("Hypothesis 3: Is there a difference in utilization rates between genders in nervous system-related cases?")
     print(gender_utilization)
     # Analyze the difference in utilization rates between genders in nervous system-related cases visually or statistically
     print("Conclusion : There is a difference in utilization rates between genders in nervous system-related cases")

     Hypothesis 3: Is there a difference in utilization rates between genders in nervous system-related cases?
     sex
     All    1.524580
     F      1.501635
     M      1.559680
     Name: util_per_1000, dtype: float64
     Conclusion : There is a difference in utilization rates between genders in nervous system-related cases
```

Fig. 4.5 Implementation of Hypothesis 3 test

The **gender_utilization** variable contains the average utilization rates for males (M), females (F), and the overall average (All). The output displays the average utilization rates per 1000 individuals for each gender.

The conclusion implies a difference in utilization rates across genders in situations related to the nervous system. In order to confirm this theory, more statistical tests or visual studies could be conducted to evaluate the relevance of the observed difference in utilization rates. The output does not specify whether the difference is higher or lower.

4.1.5 Hypothesis test 4

Here is another hypothesis that uses the one-way ANOVA (Analysis of Variance) test to determine if there are significant differences in utilization rates among different years. ANOVA evaluates the variability between groups compared to the dispersion within groups. The null hypothesis (H0) states that there are no statistically significant differences in utilization rates across different years.

The F-statistic is the quotient obtained by dividing the variation across group means by the variance within groups. A high F-statistic signifies a larger amount of variability between groups, which could potentially provide evidence for the alternative hypothesis. The p-value represents the probability of obtaining the reported data assuming that the null hypothesis (H0) is correct. A p-value that is below the standard threshold of 0.05 indicates strong evidence to reject the null hypothesis in favor of the alternative hypothesis.

```
# Perform one-way ANOVA to test for significant differences in utilization rates among years
f_stat_util, p_value_util = f_oneway(*[group['util_per_1000'] for name, group in data.groupby('yr')])

print("Hypothesis 4: Are there significant differences in utilization rates across different years?")
print(f"ANOVA F-statistic for utilization rates: {f_stat_util}")
print(f"P-value for utilization rates: {p_value_util}")
if p_value_util < 0.05:
    print("Conclusion: There are significant differences in utilization rates across different years.")
else:
    print("Conclusion: There are no significant differences in utilization rates across different years.")
```

```
Hypothesis 4: Are there significant differences in utilization rates across different years?
ANOVA F-statistic for utilization rates: 0.06733901548486773
P-value for utilization rates: 0.9772624480878692
Conclusion: There are no significant differences in utilization rates across different years.
```

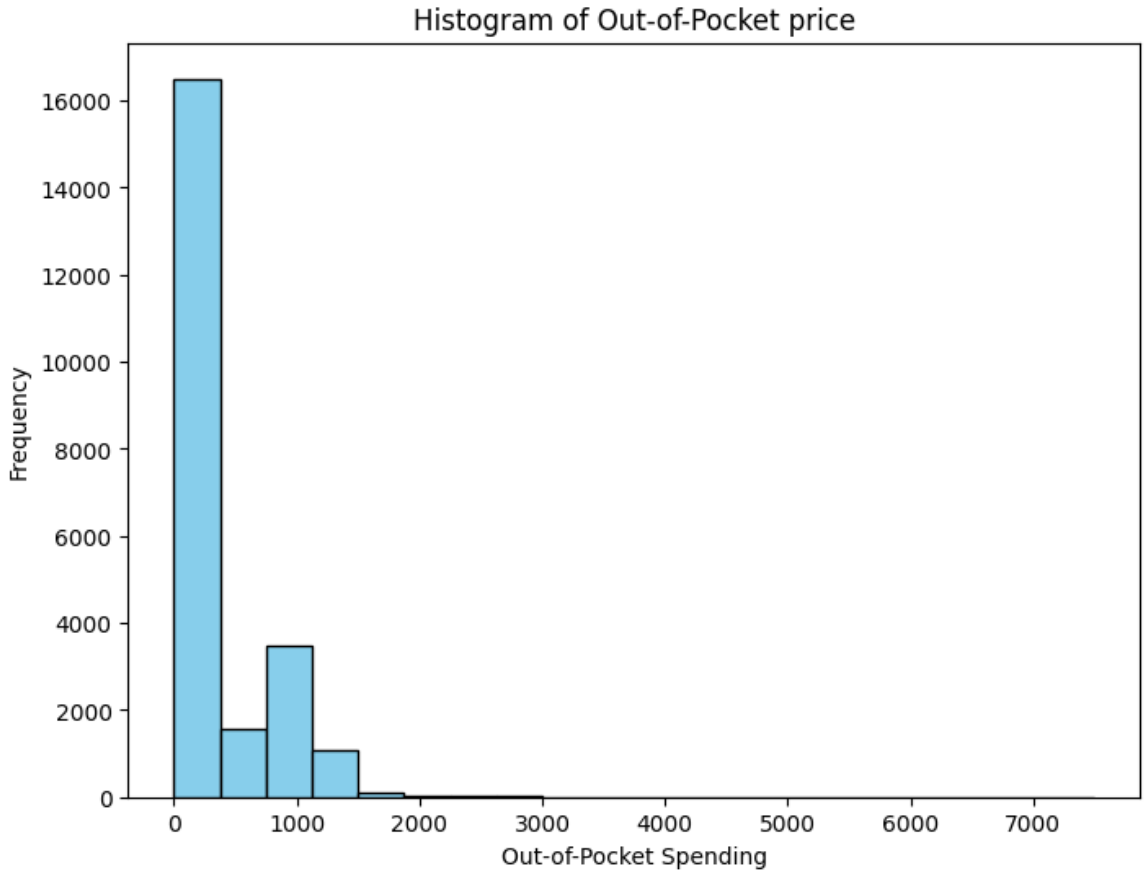
Fig. 4.6 ANOVA Implementation of Hypothesis test 4

The F-statistic in this instance is 0.0673, whereas the p-value is 0.9773. When the p-value is large, there is not enough evidence to dismiss the null hypothesis. Thus, the final hypothesis is that there are no statistically significant differences in utilization rates between various years. The findings suggest that any reported fluctuations in utilization rates are largely due to random occurrences, and no significant temporal patterns are identified in the dataset.

4.2 Data Visualization

Data visualization refers to the process of representing information and data graphically. The process entails transforming intricate datasets into graphical representations such as charts, graphs, and maps in order to enhance the accessibility and comprehension of patterns, trends, and insights. Efficient data visualization has the potential to improve the distribution of insights and facilitate the generation of outcomes.

4.2.1 Histogram



The distribution of out-of-pocket expenses among insurance holders is right-skewed in the histogram, suggesting that a smaller percentage of individuals encounter higher costs compared to the majority who encounter lower out-of-pocket spending. The range of out-of-pocket costs is denoted along the x-axis, while the frequency or count of individuals falling within each range is illustrated along the y-axis.

Within this distribution, a considerable proportion of insurance policyholders have out-of-pocket expenditures concentrated in the lower portion of the spectrum, primarily spanning from 0 to 500. This suggests that a considerable percentage of individuals experience comparatively moderate personal financial expenditures. Nevertheless, the histogram's rightward extension indicates a tail consisting of insurance holders who pay larger out-of-pocket expenses.

Comprehension of this distribution is essential in order to evaluate the economic strain experienced by members of the population under study. The right-skewed distribution indicates that although the majority of insurance holders experience manageable out-of-pocket expenses, a subset is confronted with significantly higher costs. Additional examination and investigation of this distribution may yield valuable insights regarding the determinants that contribute to the variability in out-of-pocket expenses. Moreover, it may facilitate the customization of financial strategies or healthcare policies to better cater to the distinct requirements of various demographic groups.

In order to have a deeper understanding of the histogram, we generated a histogram using the seaborn library. The histogram will display the oop_price values on the x-axis and the patient type (hcci_detcat_label_1) on the y-axis. The histogram analysis of out-of-pocket prices, categorized by patient groups, uncovers a significant pattern. More precisely, it suggests that individuals who are admitted to the hospital as inpatients generally have higher costs that they have to pay themselves, in contrast to other types of patients.

These insights enhance our understanding of the economic factors involved in healthcare usage and help in creating unique approaches to meet the different financial requirements of specific patient groups.

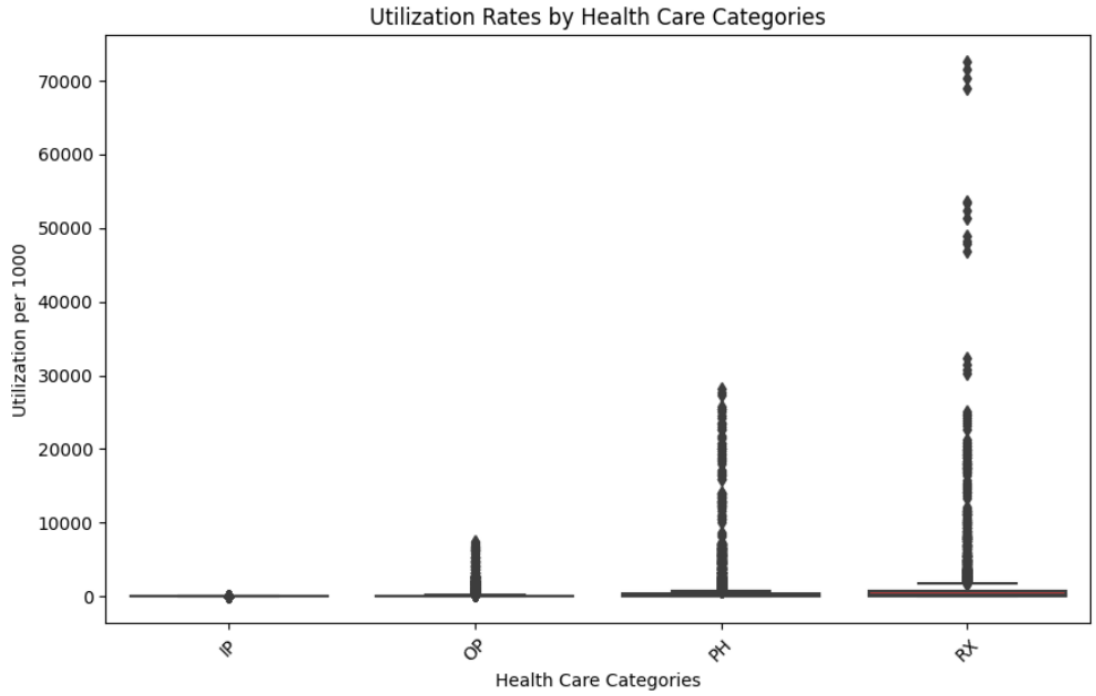
4.2.2 Barplot

The box plot analysis, with utilization on the x-axis and patient types (inpatient, outpatient, prescription drugs, and professional services) on the y-axis, provides insights into the different levels of healthcare service utilization among the population under study.

The plot reveals obvious trends in the utilization among various categories of patients. Prescription pharmaceuticals, in particular, have the greatest median utilization, indicating that a considerable number of people in the population use prescription prescriptions. This suggests a widespread dependence on pharmaceutical interventions for healthcare requirements.

Following that, professional services, such as consultations and specialist visits, are utilized, highlighting their significant contribution to overall healthcare utilization. Conversely, the utilization of inpatient services is shown to have the lowest median, indicating that fewer people need or use inpatient treatment compared to alternative healthcare methods.

Comprehending these usage patterns is essential for healthcare planning and allocation of resources. Increased usage of prescription medications may need the implementation of pharmaceutical management strategies, while decreased usage of inpatient services may inform decisions regarding hospital capacity and infrastructure.



This information provides a basis for further investigation into the factors that influence usage patterns. It allows healthcare providers and policymakers to customize services to meet the specific requirements and preferences of the population, ultimately improving the efficiency and effectiveness of healthcare delivery.

4.2.3 Line Plot

The line plot demonstrates the average price patterns from 2017 to 2021, providing useful insights into the overall cost dynamics within the healthcare dataset under study. The upward trajectory observed from 2017 to 2021 signifies a consistent rise in average prices during this timeframe.

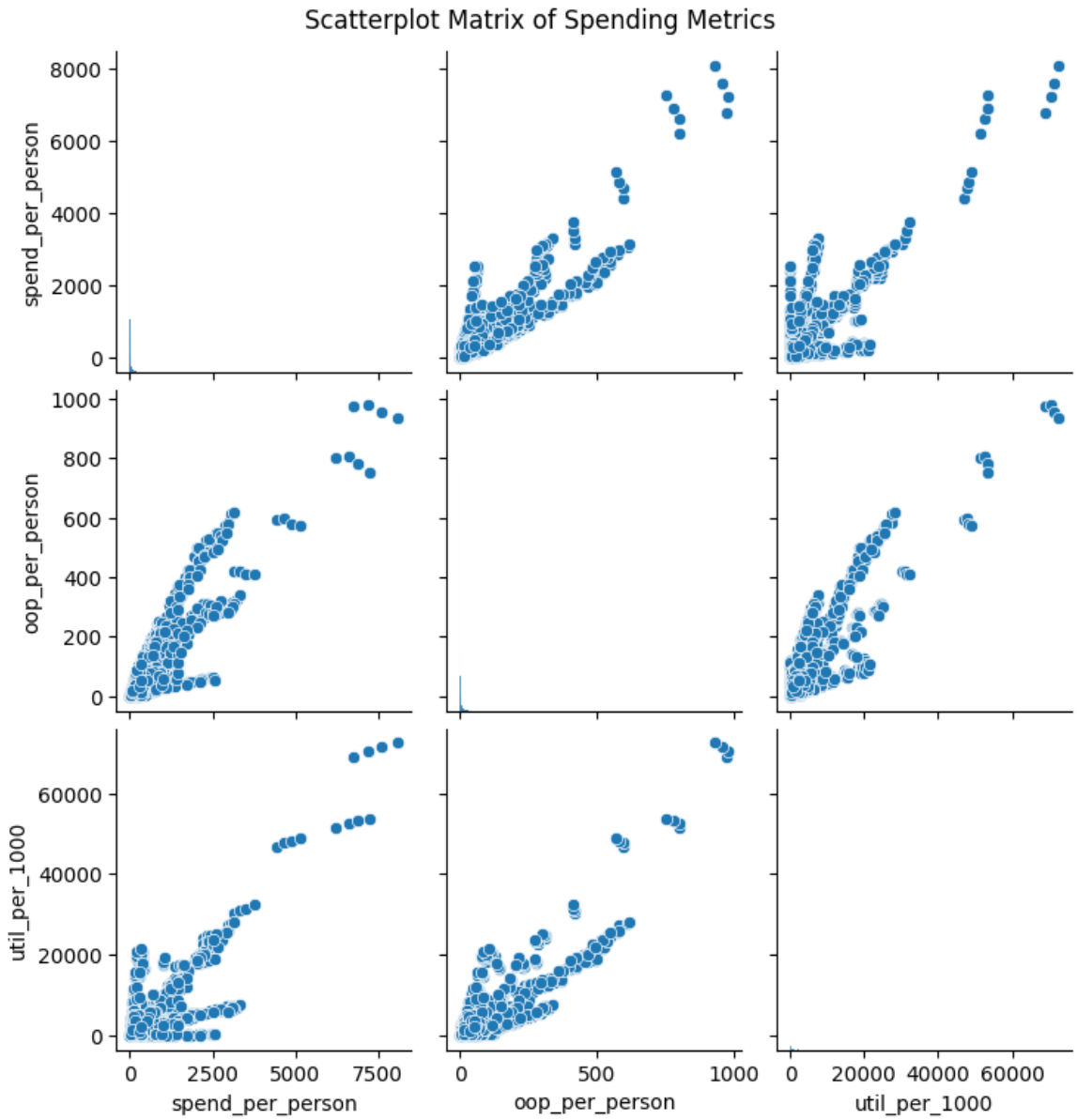
The fluctuation in average pricing over time can be affected by various factors, including inflation, developments in medical technology, or changes in patient demographics..

The increasing tendency indicates a possible increase in healthcare costs, emphasizing the necessity to conduct further research on the precise factors driving this pattern. Examining the variables that contribute to rising healthcare costs might help in devising measures to control and reduce them, thereby guaranteeing that healthcare remains affordable and accessible to the people.

Furthermore, this observation raises inquiries on the influence of these increasing expenses on patterns of healthcare utilization, results for patients, and the overall sustainability of the healthcare system. Further investigation can explore these elements in greater depth, promoting a thorough comprehension of the changing healthcare cost environment and enabling well-informed decision-making for participants in the healthcare system.

4.2.4 Scatter matrix for spending metrics

The scatter matrix, which includes spend_per_person, utilization_per_person, and out_of_pocket spent per person, is a valuable tool for identifying possible connections, patterns, and trends in the dataset. As it is evident from the figure all three of the variables are in correlation with each other. Refer to the figure on the next page.

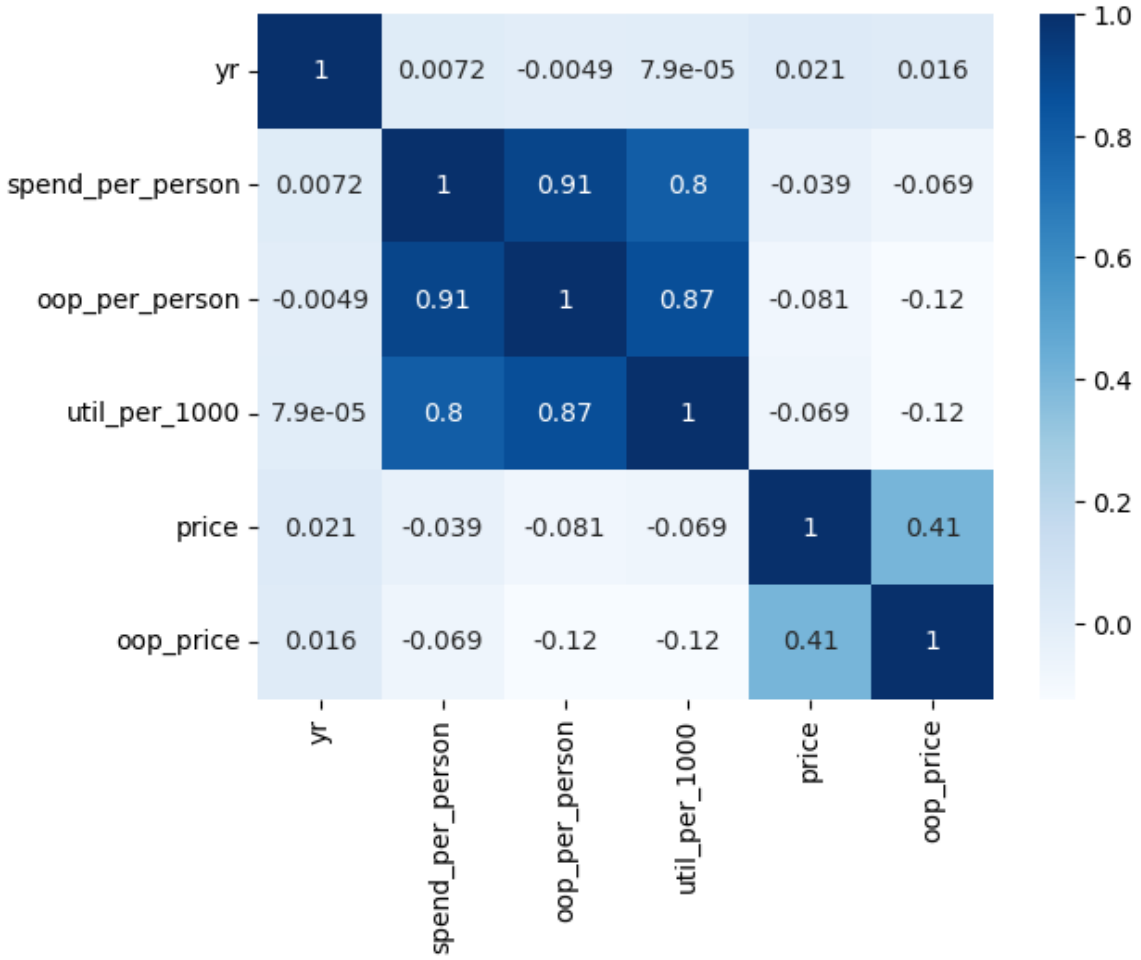


The scatter matrix reveals correlations between spend_per_person, utilization, and out_of_pocket expenditure per person, indicating correlated dynamics within the healthcare dataset. These trends can be analyzed from multiple perspectives: A scatter plot exhibiting an upward-sloping pattern indicates a positive correlation between the dots. For instance, a larger spend_per_person could be linked to an increase in utilization or higher out-of-pocket spending. We can also detect clusters or distinct patterns within data points which helps unveil subgroups within the sample. These clusters could potentially indicate separate demographic groups, health issues, or other factors that have an impact on expenditure trends.

Exceptional cases may be represented by isolated points that vary from the main trend. These outliers may suggest exceptional circumstances that require additional examination.

Gaining a comprehensive understanding of these relationships is of utmost importance for healthcare decision-makers. It enables them to precisely identify the factors that contribute to financial constraints, variances in healthcare consumption, and areas that require solutions. Analyzing these patterns might inform specific approaches to enhance the effectiveness of healthcare, decrease expenses, and optimize patient results.

4.2.5 Correlation Heatmap



The correlation Heatmap provides a visual representation of the relationships between different variables in the dataset. Here is a breif interpretation of the Heatmap:

Positive Correlations: Utilization_per_1000, spend_per_person, and oop_per_person: A positive correlation between these variables suggests that higher healthcare utilization is associated with increased spending per person. Similarly, a positive correlation indicates that higher utilization is linked to higher out-of-pocket expenses per person. The positive correlation here also implies that higher overall spending is associated with increased out-of-pocket costs per person.

Negative Correlations: Price and oop_per_person: The negative correlation between these variables suggests that as the overall price of healthcare services increases, the out-of-pocket expenses borne by individuals tend to decrease, and vice versa.

Understanding these correlations is essential for healthcare analysts and policymakers. Positive correlations can indicate areas where increased utilization might contribute to higher costs, while negative correlations may suggest complex interactions between overall prices and individual out-of-pocket burdens. This insight can guide strategies to optimize healthcare spending, improve cost-effectiveness, and enhance the financial experience for patients.

4.3 Data Preprocessing

Here we can see various crucial procedures in the data pre-processing stage, involving preparing the dataset for machine learning. Below is a comprehensive breakdown of the process:

```
# Remove null and NaN values
data = data.dropna()
```

```
# Convert 'oop_price' and 'price' columns to integer values
data['oop_price'] = data['oop_price'].astype(int)
data['price'] = data['price'].astype(int)
```

```
# List of categorical columns
categorical_cols = ['age_band_cd', 'sex', 'hcci_detcat_1', 'hcci_detcat_2', 'hcci_detcat_3',
                    'hcci_detcat_1_label', 'hcci_detcat_2_label', 'hcci_detcat_3_label']
```

Fig. 4.7 Data Preprocessing 1.1

```
# Create a ColumnTransformer to apply one-hot encoding to categorical columns
preprocessor = ColumnTransformer(
    transformers=[('cat', OneHotEncoder(), categorical_cols)],
    remainder='passthrough'  # Pass through any remaining columns
)
```

```
preprocessor
```

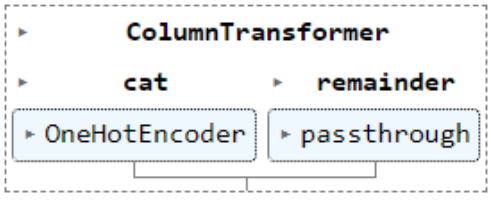


Fig. 4.8 Data Preprocessing 1.2

Handling Missing Values:

The dataset is cleansed by eliminating any Null or NaN values to maintain data integrity and mitigate potential complications during analysis.

Datatype conversion: The columns 'oop_price' and 'price' are changed to integer values, which is essential for numerical calculations and interoperability with models.

Encoding Categorical Columns: A set of categorical columns is detected, and a ColumnTransformer is generated to perform one-hot encoding on these categorical columns. It is imperative to include categorical data in machine learning models at this stage.

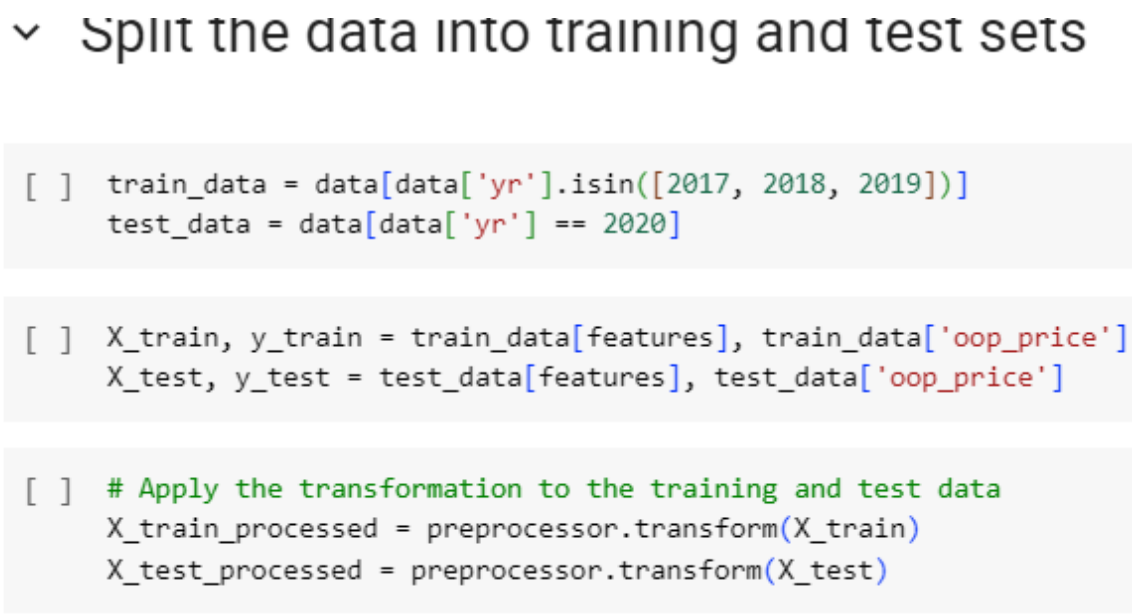


Fig. 4.9 Data Preprocessing 1.3

Data splitting: The dataset is split into training and test sets according to the year (2017-2020). It is customary to assess the model's performance on data that it has not been trained on.

The pre-processing procedures are uniformly performed to both the training and test data, guaranteeing that the same transformation is applied to different subsets of the dataset.To summarize, the code offers a thorough pre-processing pipeline that handles missing values, transforms data types, encodes categorical variables, defines features and targets, and separates the data for training and testing. These processes are essential for preparing the dataset for machine learning tasks, improving the resilience and ability of the models trained on this data to apply to different situations.

4.4 Machine Learning Models

4.4.1 Linear Regression

We applied linear regression to the HCCI dataset. For instance, it could predict out-of-pocket prices based on factors like age, gender, or specific healthcare categories. Linear regression provides insights into the linear associations between variables, helping identify the most influential factors affecting the prices. It is interpretable and can assist in understanding how changes in certain variables correlate with changes in out_of_pocket prices. However, it assumes a linear relationship, which might be a limitation if the underlying patterns are more complex.

Linear regression is a statistical technique that involves creating a mathematical model to represent the relationship between a dependent variable (oop_price) and one or more independent variables (also known as features). This is done by fitting a linear equation to the observed data. The linear regression equation, in its most basic form, consists of a single independent variable.

The equation is represented as

$$y = mx + b$$

, where m and b are constants.

Here, the dependent variable is represented by y, the independent variable is represented by x, the slope of the line is denoted by m, and the y-intercept is denoted by b. For situations involving numerous independent variables, the equation is expanded to:

The expression can be simplified as

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

.

The y-intercept is represented by b_0, while the coefficients for each independent variable are represented by

$$b_1, b_2, ..., b_n$$

Linear Regression Evaluation: The given case fits a linear regression model to the preprocessed training data and assesses its performance on the test set. The intercept of about -9.18 billion is the expected out-of-pocket expenditure when all independent variables have a value of zero.

The Mean Squared Error (MSE) quantifies the average of the squared differences between the predicted and actual values. On the other hand, the R-squared score indicates the

proportion of the variability in the dependent variable that can be anticipated based on the independent variables. A higher R-squared value, which is closer to 1, suggests a more optimal fit of the model.

Significance of Findings: The Mean Squared Error, which is around 38,970, and the R-squared score of 0.81 indicate that the linear regression model effectively accounts for a significant amount of the variation in out-of-pocket spending. The R-squared score suggests that around 81% of the variance in out-of-pocket spending can be explained by the features employed in the model.

Gaining insight into the correlation between characteristics and medical expenses is essential for maximizing the distribution of resources, formulating policies, and delivering tailored healthcare. Within this particular framework, the linear regression model offers valuable insights into the extent to which alterations in particular variables contribute to fluctuations in out-of-pocket spending. Policymakers and healthcare professionals can employ these observations to discern the aspects that impact costs and execute focused interventions to effectively manage and diminish healthcare expenditures for both individuals and the healthcare system at large.

4.4.2 Random Forest

Random Forest is a powerful ensemble learning algorithm that can be beneficial for this project. It can handle both regression and classification tasks effectively. Random Forest excels in capturing complex, non-linear relationships and interactions between various features. In the context of oop_price analysis, it can provide accurate predictions by considering the importance of different features. Moreover, it is robust against overfitting and capable of handling large datasets with numerous variables. Random Forest can be particularly useful for identifying the key drivers of healthcare costs and making predictions with high accuracy. It also offers insights into feature importance, aiding in the understanding of which variables contribute most to cost variations. Random Forest is a popular ensemble learning method that is extensively employed for both regression and classification tasks. During the training process, this system creates several decision trees. When given an input, it produces the average prediction (for regression) or the most common prediction (for classification) from the individual trees.

The general formula for a single decision tree within the context of random forest regression is:

$$y = \frac{1}{N} \sum_{i=1}^{N} h_i(x)$$

The equation represents the sum of the values of a function h(x) multiplied by the corresponding weights yi, where i ranges from 1 to N. The result is denoted by y

What is the location or place where something is happening or being referred to?

The anticipated output, y, is the result that is expected. N represents the number of trees in the random forest. $h_i(x)$ denotes the forecast made by the i_th tree. Within the scope of your project, the Random Forest algorithm amalgamates forecasts from numerous decision trees to yield a resilient and precise prediction of out-of-pocket expenses. The collective structure of Random Forest aids in reducing overfitting and capturing intricate correlations within the data.

Regarding the project's precise discoveries:

The Mean Squared Error (MSE) and R-squared Score are computed to evaluate the effectiveness of the Random Forest regression model. The Mean Squared Error (MSE) quantifies the average of the squared differences between anticipated and actual values, where smaller values indicate superior performance. The R-squared Score measures the amount of variation in the dependent variable that is explained by the model, with a value of 1.0 indicating a perfect fit. When considering the results within the scope of your project, a high R-squared Score and a low MSE suggest that the Random Forest model is proficient at forecasting out-of-pocket expenses. This offers significant insights for analyzing healthcare costs.

4.4.3 Support Vector Regression (SVR)

The Support Vector Regressor(SVR) is a regression algorithm that extends support vector machines to predict continuous outcomes. In this project, The Support Vector Regressor (SVR) is applied to model the relationship between independent variables and out-of-pocket prices. SVR is effective in capturing complex patterns in the data and is less sensitive to outliers. It works well in high-dimensional spaces, making it suitable for datasets with multiple atfeures. The Support Vector Regressor (SVR) is specifically designed for regression tasks. SVR, in comparison with linear regression, is a non-linear model that effectively handles complex patterns within the data. Below is a comprehensive clarification, mathematical equation, and analysis of Support Vector Regression (SVR):

Support Vector Regression (SVR) seeks to identify a hyperplane that maximizes the margin while allowing for a certain amount of error. This process effectively constructs a tube around the regression line. Support vectors, which are data points located near the margin, are utilized to establish the tube. The regression line precisely travels through the center of the tube.

$$haty = \sum_{i=1}^{N} (\alpha_i \cdot K(x, x_i)) + b$$

Let N be the number of support vectors α_i refers to the Lagrange multipliers denotes a specific point. The kernel function

$$K(x, x_i)$$

quantifies the similarity between x and the support vector x_i. The bias term is denoted as b.

Explanation: The Mean Squared Error (MSE) is a statistical measure that calculates the average of the squared differences between predicted and actual values. This metric quantifies the mean squared deviation between expected and actual values. A smaller mean squared error (MSE) signifies superior predicting ability. The R-squared score (R^2) quantifies the fraction of the variance in the dependent variable that can be accurately predicted by the independent variables. A number closer to 1 implies a higher level of accuracy in fitting the data. In our case, the Mean Squared Error (MSE) and R-squared (R^2) scores for Support Vector Regression (SVR) should be evaluated in the same manner as those for linear regression and random forest. A decrease in Mean Squared Error (MSE) and an increase in the coefficient of determination (R^2) indicate that the Support Vector Regression (SVR) model effectively captures the underlying patterns in the data. But here the R^2 score of 0.64 suggests that the SVR model accounts for approximately 64% of the variance in the out-of-pocket prices, indicating a moderate level of explanatory power.

In summary, Linear Regression provides simplicity and interpretability, Random Forest excels in capturing complex relationships, and SVR handles non-linear patterns effectively. The selection among these models depends on the specific characteristics of the data and the underlying relationships in the healthcare cost domain. A combination of these models or an ensemble approach may also be explored for improved predictive performance.

4.5 Neural Networks

Compilation: The model is compiled with the Adam optimizer and the mean squared error (MSE) loss function. Adam is a widely used optimization technique, and Mean Squared Error (MSE) is appropriate for regression issues involving the prediction of a continuous variable.

Training Epochs: The model undergoes 10 epochs, which entails processing the complete dataset 10 times. During each epoch, the model adjusts its weights in order to minimize the training loss.

Loss Reduction: The mean squared error (MSE) decreases during epochs, suggesting that the model is acquiring knowledge and enhancing its capability to generate predictions. The training and validation losses, which are presented for each epoch, exhibit a declining trend, indicating that the model is not excessively fitting to the training data.

Validation Loss: The validation loss is an essential parameter used to evaluate the performance of a model during training. It quantifies the model's ability to generalize to unseen data. The declining validation loss suggests that the model is not merely memorizing the training data, but also acquiring knowledge of patterns that are applicable to unseen data.

Epoch Duration: The duration of each epoch is shown. Training smaller datasets with fewer features typically results in faster training times, whereas the duration can be influenced by the intricacy of the neural network structure.

Generally, the declining loss numbers indicate that the neural network is acquiring knowledge from the data. Following the completion of training, the model can be utilized to provide predictions on novel and unobserved data. Hence, we decided to assess the model's performance on a **validation test set (2021)** to ensure that it demonstrates good generalization and avoids overfitting to the training data. We will study the evaluation of different models in the next Chapter

Chapter 5

Evaluation

5.1 Model Evaluation

When assessing the effectiveness of various regression models, such as Linear Regression, Random Forest (RF), and Support Vector Regressor (SVR), you can employ a range of metrics and visualization tools. Below is a comprehensive and systematic way of comparing your models and evaluating your outputs.

1. Assessment criteria:

The Mean Squared Error (MSE) is a metric that quantifies the average of the squared differences between expected and actual values. A lower mean squared error (MSE) suggests superior performance. The R-squared score, often known as R2, quantifies the amount of the variance in the dependent variable that can be accurately predicted. A greater coefficient of determination (R2) indicates a more accurate and precise fit between the observed data and the regression model.

2. Scatter Graph:

The X-axis represents the real values or ground truth. The Y-axis represents the predicted values generated by the respective models. Every point on the scatter plot corresponds to a data point in your test set.

3. Evaluation Process:

Compute the Mean Squared Error (MSE) and R-squared (R2) for each model utilizing the test set. Analyze the metrics to ascertain the superior performance of each model.

4. Analysis:

Linear Regression: Linear Regression has a Mean Squared Error of 38969.91840610617 R-squared Score of 0.8107360126858644. This means Linear Regression is working moderately better than The Support Vector Regressor.

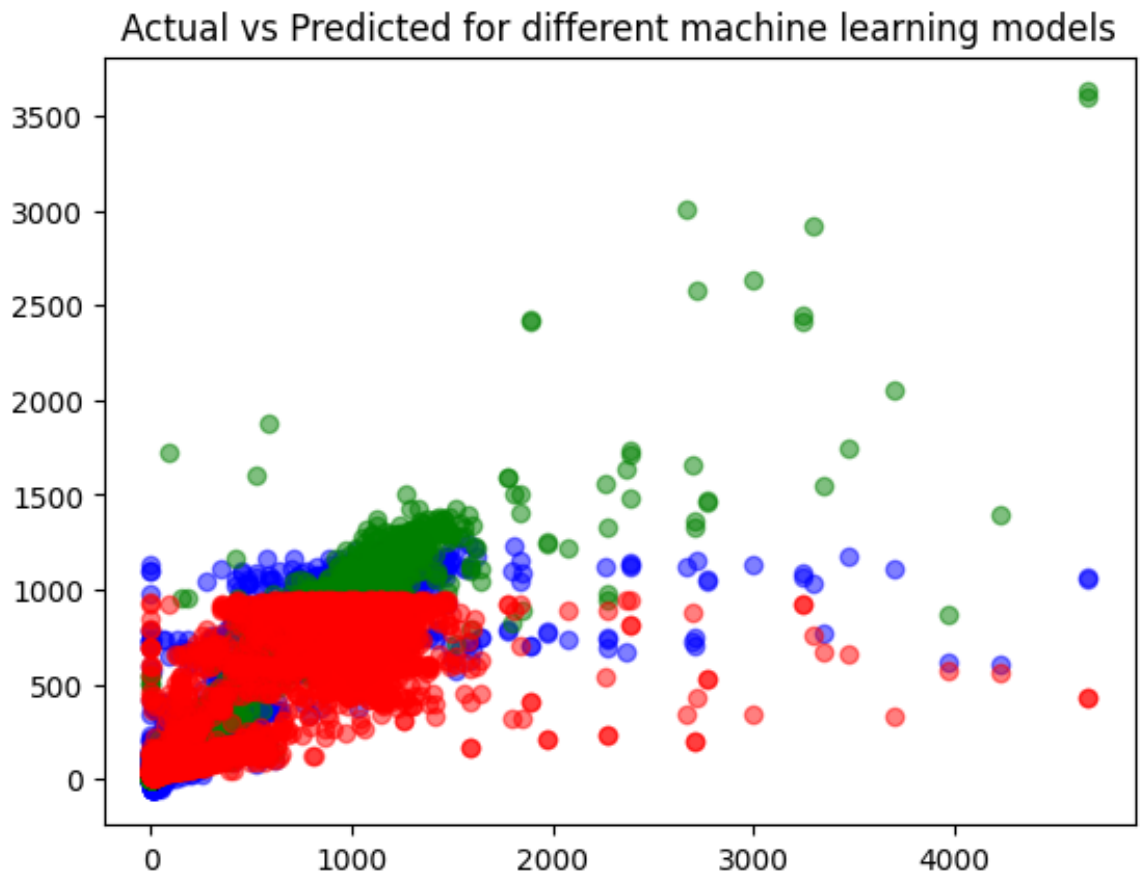


Fig. 5.1 Model Evaluation 1.1

Random Forest: Random forrest has a Mean Squared Error of 14088.281094496055 and R-squared Score of 0.9315778845990891, it suggests that the model performs well in capturing non-linear patterns and interactions. Also, the best performance by any model so far.

The Support Vector Regressor (SVR) can effectively capture patterns without overfitting when it achieves a balance between Mean Squared Error (MSE) and R-squared (R2).But in our case, SVR has Mean Squared Error of 74654.68585473597 and R-squared Score of 0.637426915568353 making it the least effective model amongst the other two.

Optimal situation: The points on the scatter graph should align in a diagonal line, suggesting a close correspondence between predicted and actual values.

In conclusion:

The Random Forest (RF) algorithm has superior performance compared to both Linear Regression and Support Vector Regressor (SVR), with the highest R-squared score of 0.93. The Linear Regression model demonstrates a commendable R2 value of 0.81, which indicates

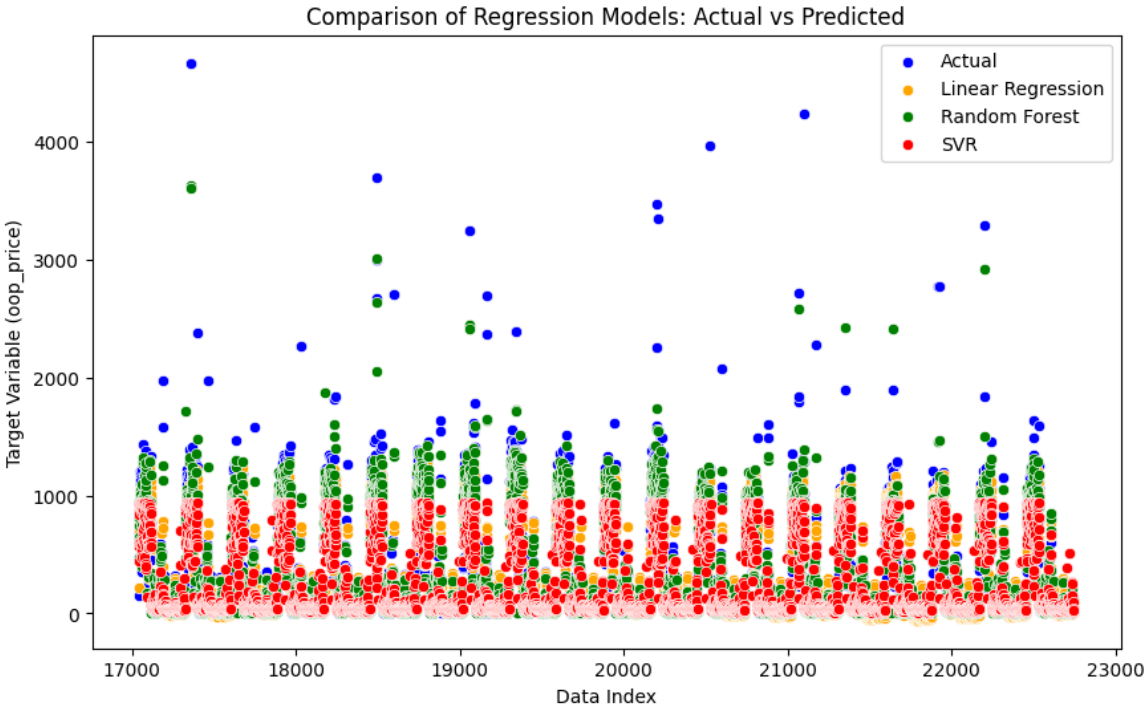


Fig. 5.2 Neural Network Evaluation 1.2

a strong predictive power. The Support Vector Regressor (SVR) exhibits a lower R2 value of 0.63, indicating a somewhat inferior predictive capability.

Observations: The improved performance of Random Forest can be attributed to its capacity to capture non-linear correlations and interactions within the data. Linear Regression offers a robust and easily understandable model, but SVR, however having a lower R2, may face difficulties in capturing the intricacy of the underlying patterns. When selecting a model for your project, it is important to carefully consider the application requirements and the model's interpretability. This conclusion offers a brief overview of the comparative performance of the three models, as determined by their R-squared values. Nevertheless, it is essential to incorporate the Mean Squared Error (MSE) values and maybe other pertinent metrics to conduct a thorough evaluation.

5.2 Neural Network Evaluation

Model Assessment using Validation Set (2021): A neural network model trained on prior data is assessed on the validation set (2021 data) to measure its performance on unseen data.

Assessment criteria:

The Mean Squared Error (MSE) is 59821.03 and the R-squared Score (R2) is 0.77. DataFrame containing the results of the test:

Two DataFrames, test_results and eval_results, are created to store the actual and predicted values for the test set (2020 data) and the validation set (2021 data), respectively. Test

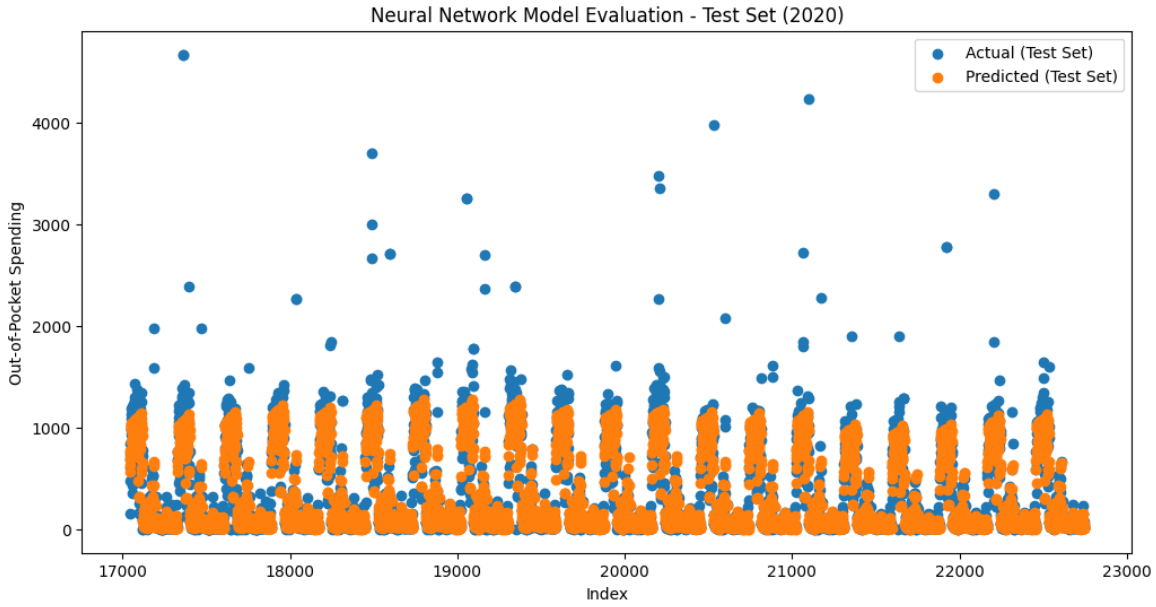


Fig. 5.3 Model Evaluation 1.2

Set Scatter Plot (2020):

A scatter plot is created to visually compare the real and expected out-of-pocket spending amounts for the test set in 2020. Every point on the graph corresponds to an index, with the y-axis indicating the amount of out-of-pocket expenses.

Assessment:

The neural network model has satisfactory performance on the validation set, as evidenced by the R-squared score of 0.77. This indicates that the model effectively captures a significant amount of the variation in the target variable.

Nevertheless, the Mean Squared Error (MSE) value of 59821.03 suggests the presence of prediction inaccuracies. It is essential to analyze the Mean Squared Error (MSE) considering the particular project needs and the magnitude of the target variable.

Visual examination:

The Scatter plot provides a visual comparison between the actual values and the expected values for the test set in the year 2020. The closeness of the two points demonstrates the model's capacity to replicate the real spending pattern. Clutters observed in the lines can indicate the specific regions where the model faces difficulties or demonstrates exceptional performance in making predictions.

Suggestion:

Additional analysis, such as evaluating the relevance of features or optimizing hyperparameters, has the potential to enhance the model's performance. Examine the mispredictions to improve the model's abilities. To summarize, although the neural network model exhibits potential, a comprehensive evaluation necessitates a profound comprehension of the Mean Squared Error (MSE) and a visual examination of predictions. Modifications or improvements to the model can be implemented according to the precise objectives and specifications of your project.

Further, We generate two additional scatter plots to visually represent and compare the real and predicted out-of-pocket expenditure amounts for the evaluation set (2021 data), as well as to compare predictions between the test set (2020) and the evaluation set.

Line plot depicting the evaluation set using data from 2021:

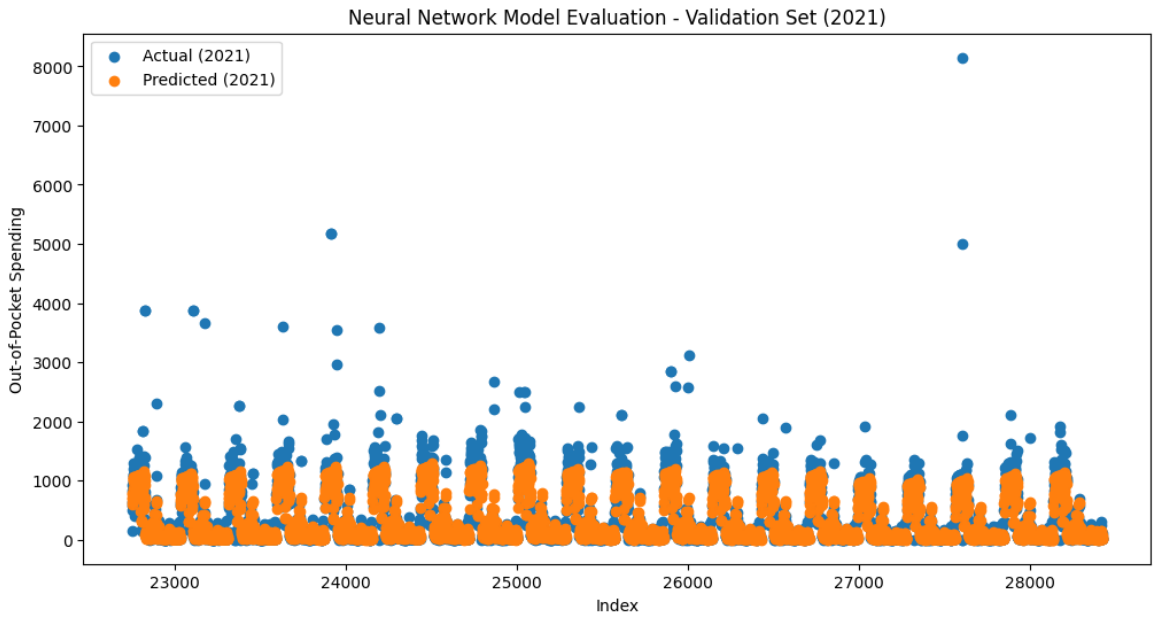


Fig. 5.4 Neural Network evaluation 1.2

This scatter plot illustrates the comparison between the real and forecasted out-of-pocket expenditure figures for the evaluation set in the year 2021. Every data point on the graph

corresponds to an index value, while the vertical axis shows the amount of money spent directly by individuals. The plot facilitates the evaluation of the neural network model's performance on novel and unobserved data from the year 2021. Scatter plot that combines data from both the test and evaluation sets:

The scatter plots for the test set (2020) and evaluation set (2021) are merged into a unified plot. The actual and anticipated values from both sets are visually compared by examining the points. The integration of this graphic enables a thorough evaluation of the model's performance across various time intervals. Explanation:

The scatter plot of the evaluation set (2021) facilitates the assessment of the model's capacity to generalize to novel data. The closeness of the actual and projected points offers valuable information regarding the model's precision on this particular group of data. The merged scatter plot enables a straightforward visual comparison between predictions

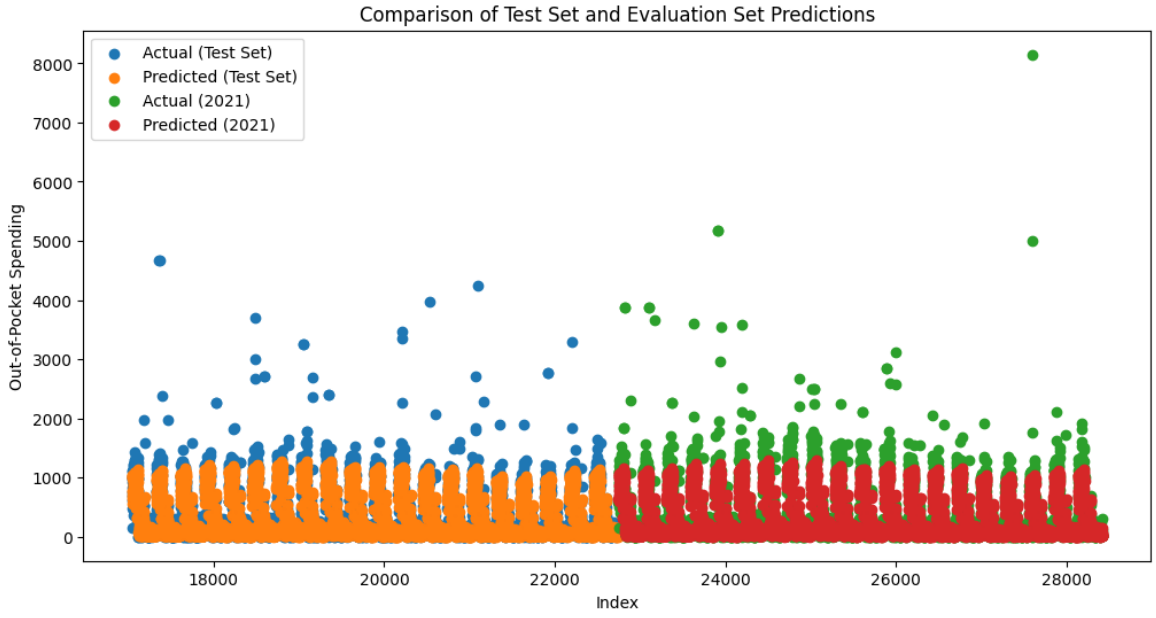


Fig. 5.5 Caption

made on the test set (2020) and the evaluation set (2021). Evaluating the degree of concordance between the model's forecasts and the real expenditure in both datasets facilitates comprehension of the model's durability and resilience over time.

Suggestion:

Examine the congruity between the real and projected data points in both scatter diagrams. The presence of consistent patterns over many time periods suggests a model that exhibits strong generalization capabilities. These visuals enhance the quantitative measurements (MSE and R-squared) previously presented, providing a comprehensive comprehension of the

neural network model's performance across various datasets and time periods. Modifications or enhancements to the model can be contemplated based on the observations obtained from these visual evaluations.

5.3 Evaluation Summary

The evaluation of three predictive models - Linear Regression, Random Forest, and Support Vector Regressor (SVR) - revealed that Random Forest is the most accurate model for predicting out-of-pocket spending. Linear Regression, with an R-squared score of 0.81, explains 81% of the variance in out-of-pocket spending. Random Forest, with a lower MSE and high R-squared score, captures 93% of the variability. SVR, with an R-squared score of 0.64, explains 64% of the variance. The choice of model depends on specific requirements, interpretability, and computational considerations. The results provide insights for informed decision-making, resource optimization, and strategic planning in healthcare financing. After which The Neural Network model was evaluated using Mean Squared Error (MSE) and R-squared Score, focusing on its predictive performance on the validation set (2021 data). The model showed reasonable predictive performance, explaining 77% of the variability in out-of-pocket spending. However, the Random Forest model outperformed the Neural Network model in terms of both MSE and R-squared score. The choice between these models depends on specific requirements and priorities, considering factors like interpretability, computational complexity, and the trade-off between accuracy and explainability. Scatter plots were created to visualize the model's performance. Accurate predictions from these models help healthcare providers, insurers, and policymakers understand and anticipate spending trends, optimizing resource allocation and financial planning.

Chapter 6

Conclusion and Suggestions

6.1 Conclusion

This study explores out-of-pocket healthcare expenditure, focusing on demographics, out-of-pocket spending patterns, and healthcare utilization. Using statistical techniques, hypothesis testing, and ANOVA, the study found significant differences among variables. Machine learning models, including Linear Regression, Random Forest, Support Vector Regressor, and Neural Network, were used to analyze the data. Key findings revealed variations in spending across age groups, genders, and health categories.

Utilization rates revealed trends, particularly in nervous system-related cases, highlighting gender disparities. The study also established statistically significant relationships and predictive capabilities within the dataset. The Random Forest model was found to be the most effective, providing a balance between accuracy and interpretability. The findings can be used by stakeholders like healthcare providers, insurers, and policymakers for informed decision-making. However, the study acknowledges the complexities of healthcare economics and the potential influence of unmeasured variables. Future research could explore specific healthcare categories, temporal trends, or external factors influencing out-of-pocket expenditure.

6.2 Suggestions

1. **Improved Feature Engineering**: executing a thorough analysis to include additional variables can enhance the predictive models. To reach a comprehensive set of elements, one must broaden the understanding of healthcare utilization patterns together with statistical, territorial, and financial aspects. Considering aspects such as healthcare methodology, prevalent diseases, or lifestyle choices may potentially enhance predictive accuracy.

2. **Temporal Analysis and Trend Identification**: By conducting a thorough analysis of temporary changes and identifying recurring patterns in healthcare expenses, it appears that the accuracy of predictions can be improved. Models that incorporate dynamic patterns occurring throughout many periods, including annual, seasonal, or monthly, have the ability to identify subtle fluctuations and improve the accuracy of predictions.

3. **Ensemble Modeling Techniques**: Employing ensemble tactics, such as stacking or combining predictions from several models, appear to leverage the strengths of individual computations. Merging the outputs of various models can mitigate biases and enhance predictive accuracy by integrating their diverse perspectives.

4. **Dynamic Data Integration**: The continuous integration of enhanced datasets and real-time information streams can maintain relevance based on demonstrated usefulness. Acquiring dynamic data appears to provide aid in quickly adapting to evolving healthcare factors and demonstrating effects, fostering more adaptable predictive models.

5. **EDA (Exploratory Data Analysis):** A thorough exploratory data analysis (EDA) appears to uncover concealed linkages and patterns within the dataset. Advanced visualizations or relationship analyses may uncover dormant connections between components, enhancing the process of determining highlights and conveying insights.

6. **Enhancement in Evaluation Metrics**: Incorporating alternative evaluation metrics beyond traditional measures (MSE, R-squared) may offer a more comprehensive perspective on performance evaluation. Metrics such as Mean Absolute Error (MAE) or quantile regression evaluation appear to provide a more comprehensive knowledge of predictive accuracy across many contexts.

7. **Ethical Considerations and Bias Mitigation**: It is crucial to prioritize moral deliberations and the reduction of bias. Implementing standard reviews and conducting decency checks in predictive algorithms helps ensure fair distribution of healthcare resources, preventing biases against specific socioeconomic groups or regions.

8. **Cooperation and cross-disciplinary perspectives:** Engaging in collaboration across several domains such as healthcare, information science, and policy-making can lead to fascinating opportunities. The integration of several areas of expertise can result in innovative strategies and comprehensive plans for accurate healthcare consumption predictions.

References

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730.

Chen, I. Y., Joshi, S., and Ghassemi, M. (2020). Treating health disparities with artificial intelligence. *Nature medicine*, 26(1):16–17.

Chen, J., Vargas-Bustamante, A., Mortensen, K., and Ortega, A. N. (2016). Racial and ethnic disparities in health care access and utilization under the affordable care act. *Medical care*, 54(2):140.

Finkelstein, A., Zhou, A., Taubman, S., and Doyle, J. (2020). Health care hotspotting—a randomized, controlled trial. *New England Journal of Medicine*, 382(2):152–162.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216.

O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Song, J., Gao, Y., Yin, P., Li, Y., Li, Y., Zhang, J., Su, Q., Fu, X., and Pi, H. (2021). The random forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms. *Risk Management and Healthcare Policy*, pages 1175–1187.