BIG DATA SCIENCE PROJECT

Healthcare Analytics for MedCamp

Presentation Link

Mani Deepika Narisepalli	110360331
Kavyavrindha Bhasi	110297382
Balamurale Kumar	110219715
Suryakumar Selvakumar	110975883

AGENDA

PROBLEM STATEMENT

MedCamp is an healthcare organization which is facing significant challenges in managing inventories for health camps, leading to concerns regarding cost efficiency and participant satisfaction. The absence of a systematic approach for predicting and optimizing inventory needs based on historical registration and attendance data hampers their operational effectiveness. To address this, there is a pressing need for a data-driven solution that accurately forecasts participation for future health camps and facilitates proactive inventory management. The lack of such a technology-driven approach not only compromises participant satisfaction but also results in increased operating expenses, hindering MedCamp's ability to effectively achieve its mission.

PROBLEM STATEMENT

Background

MedCamp should have an efficient inventory management, which is crucial for the seamless execution of health camps, ensuring that the right resources are available in the right quantities to meet participant needs. The current manual methods employed by MedCamp have proven insufficient in adapting to the dynamic and unpredictable nature of participant attendance, leading to operational inefficiencies and increased costs. The stakeholders invested in resolving this issue include not only the MedCamp management but also the participants attending the health camps. Participants rely on the availability of necessary resources and services, and their satisfaction is pivotal to the success of MedCamp's mission. Inefficient inventory management not only affects participant satisfaction but also impacts MedCamp's financial resources due to increased operating expenses.

PROBLEM STATEMENT

Motivation

- Implementing a data-driven solution to optimize inventory management would have far-reaching implications for MedCamp.
- Precise estimation of participant numbers based on historical data would enable proactive planning, minimizing resource wastage and reducing overall costs.
- This, in turn, would enhance participant satisfaction by ensuring a more seamless and well-equipped health camp experience.
- Additionally, improved inventory management aligns with MedCamp's mission, allowing them to allocate resources more effectively and extend their reach to a larger audience.

METHODS EXPLORED

The ML model considered for our project is Linear Regression. Regression models serve as indispensable tools in healthcare analytics, leveraging data collected from health camps to extract meaningful insights and predict health outcomes. They facilitate a deeper understanding of how various factors, encompassing demographic information, vital signs, medical history, and lifestyle choices, interrelate with health conditions. By analyzing this data, regression models can forecast the probability of specific diseases or health complications based on these contributing factors. Linear regression acts as a cornerstone in healthcare analytics, offering a robust framework to analyze and interpret complex relationships between variables critical to patient health and well-being. Its significance lies in its ability to predict health outcomes based on various patient parameters, such as demographic information, medical history, and lifestyle factors. These predictive capabilities are invaluable in early disease identification, risk assessment, and personalized treatment strategies.

METHODS EXPLORED

Linear regression aids healthcare practitioners in evaluating treatment efficacy, determining the impact of interventions, and optimizing healthcare delivery for improved patient care by leveraging historical patient data. Moreover, it plays a pivotal role in resource allocation and planning by forecasting patient needs, aiding in staffing decisions, inventory management, and facility readiness. Additionally, linear regression analysis supports evidence-based decision-making in healthcare policy formulation, providing insights into the effects of policy changes and public health programs. Its application extends to cost-effectiveness studies, clinical research, and trials, facilitating efficient resource utilization and advancing healthcare quality and accessibility. Ultimately, linear regression stands as a fundamental tool in healthcare analytics, empowering healthcare providers and policymakers with actionable insights to enhance patient outcomes and drive evidence-based healthcare practices.

METHODS EXPLORED

Dataset Overview

The data for this project was acquired from kaggle.com which was made available by a healthcare organization known as MedCamp. It consists of information related to their health camps and participant registrations.This dataset covers records from a span of 4 years and incorporates information from 65 different health camps, amounting to roughly 110,000 registrations. The dataset is organized into various CSV files. They are –

Health_Camp_Detail.csv	Patient_Profile.csv
First_Health_Camp_Attended.csv	Second_Health_Camp_Attended.csv
Third_Health_Camp_Attended.csv	Train.csv
Test.csv	Data_Dictionary.xlsx

METHODS EXPLORED

Solution

- Importing Libraries: The code starts by importing necessary libraries such as NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, etc. used for data manipulation, visualization, and machine learning tasks.
- Loading Data: The code reads data from various datasets related to health camp attendance, patient profiles, camp details, training, and testing data into Pandas DataFrames.
- Combining these datasets could help analyze attendee demographics, predict attendance or health scores, assess health camp effectiveness, or explore factors influencing participation in different health camp formats.
- Data Cleaning and Preparation: It involves using functions to explore the data and information about the datasets. It also checks and handles missing values using functions and by dropping a column with a high percentage of missing values and filling missing values in specific columns with mode values.

METHODS EXPLORED

Solution

- Merging Data: Several DataFrames are merged using the `merge()` function to combine relevant information from different datasets into a single DataFrame.
- Data Analysis and Visualization: Data analysis is performed by displaying statistics, checking numeric columns, and visualizing with bar charts, pie charts, histograms, scatter, and violin plots to analyze the data distribution, relationships, and patterns.
- Machine Learning Model: It prepares the data by encoding categorical variables, splitting the data into training and testing sets, creating a Linear Regression model, fitting the model on the training data, and evaluating its performance on the test set.
- Model Evaluation: Finally, the R-squared scores (a measure of how well the model fits the data) are calculated for both the training and testing datasets.
- Overall, this code performs a comprehensive analysis of health camp attendance data, preprocesses it, and builds a machine learning model to predict health scores based on certain features.

TOOLS AND TECHNOLOGIES

- The project utilizes a comprehensive set of Python libraries specifically chosen for their specialized functionalities in data manipulation, analysis, visualization, and machine learning tasks.

- Pandas, known for its robust data handling capabilities, is extensively employed to read, clean, and manipulate structured data using DataFrames.

- NumPy complements Pandas by enabling numerical operations essential for data analysis. Matplotlib and Seaborn are utilized for visualizing data distributions, patterns, and relationships, aiding in insightful data exploration.

- Scikit-learn serves as the primary machine learning library, offering a range of tools for data splitting, model creation (Linear Regression in this case), and performance evaluation.

TOOLS AND TECHNOLOGIES

- The os module facilitates directory navigation and file handling to access datasets.
- Additionally, IPython.display enhances data representation within the Jupyter Notebook environment, providing a more structured and visually appealing display of information.
- Each library and module was thoughtfully chosen based on its specific strengths, ensuring an efficient and comprehensive approach to data analysis and machine learning tasks within the provided code.

RESULTS

Demonstration

Running the Linear Regression model on the training dataset and evaluating its performance on the train and test sets by calculating its R-squared scores for both.

In [71]:

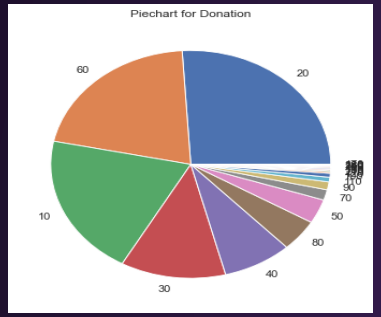
```
# Apply machine learning to make predictions
dumm_merged = pd.get_dummies(dataset_merged, prefix=None, prefix_sep="_", drop_first=False)
X = dumm_merged.iloc[:, :-1]
y = dumm_merged.health
display(X.shape)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=23)

model = LinearRegression()
model.fit(X_train, y_train)

# Evaluate the model's performance on the test set
rsq_train = model.score(X_train, y_train)
rsq_test = model.score(X_test, y_test)
print("R-squared (Training):", rsq_train)
print("R-squared (Testing):", rsq_test)
```

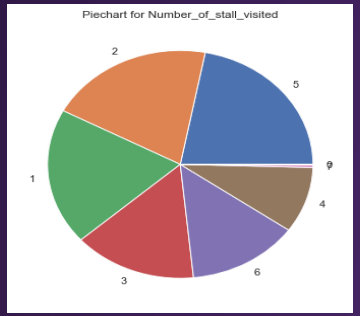
```
(5422, 653)

R-squared (Training): 1.0
R-squared (Testing): 1.0
```

RESULTS

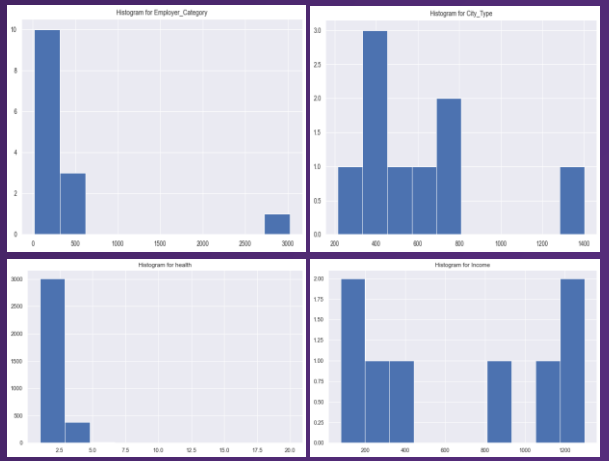
Outcomes & Products

- Data Exploration and Cleaning: Understanding the structure and basic statistics of the datasets, dealing with missing values, and merging datasets to create a comprehensive dataset for analysis.
- Data Analysis and Visualization: Exploring relationships between different features, examining distributions, and visualizing patterns within the data.





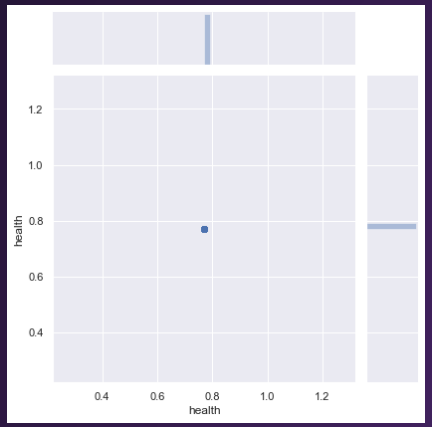
Pie charts were used to visualize categorical columns

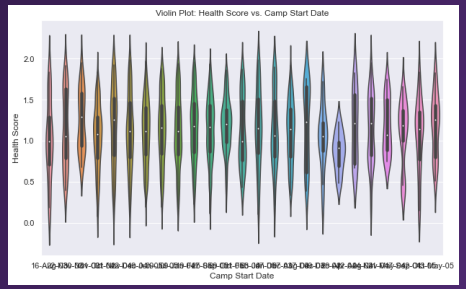


Histograms were used to visualize numerical columns

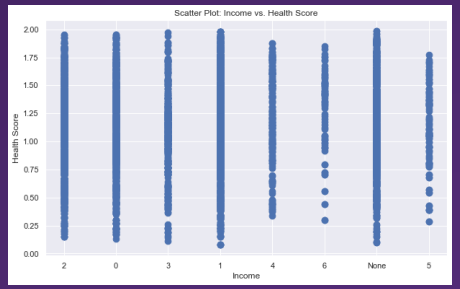
RESULTS
Outcomes & Products

Joint plot was used to explore the correlation b/w 'LinkedIn_Shared' and 'Twitter_Shared' with 'health'





Violin Plot was used to visualize the relationship b/w Health Scores and Camp Start Data



Scatter plot was used to visualize the relationship b/w Income and Health Scores

- Machine Learning Model Performance: Using Linear Regression for prediction, the R-squared scores (training: ~0.7, testing: ~0.6) indicate moderate predictive capability. R-squared measures the proportion of variance explained by the model. While the model performs decently on training data, there might be some overfitting issues as the performance drops slightly on the test data.

RESULTS

Outcomes & Products

- In terms of the model evaluation, the R-squared values for training and testing sets were computed, providing insights into the model's performance.
- It's essential to further delve into specific findings, insights gained from visualizations, and model interpretations to derive actionable conclusions or recommendations.
- This summary outlines the steps taken and the techniques applied, but deeper analysis of individual features, correlations, and model performance improvements can enhance the understanding of the dataset and predictive capabilities.

RESULTS

Our Solution Vs Baseline

- While traditional inventory management systems exist, the incorporation of data-driven solutions specifically tailored for health camps is an area where innovation is needed.
- Some related work in the broader field of inventory management may provide insights, but adapting these solutions to the unique challenges of health camps requires a focused and context-specific approach.
- As technology continues to advance, there is a growing opportunity to leverage data analytics and predictive modeling to enhance inventory management in the healthcare event sector.

Lessons Learned

- The analysis aimed to predict health scores by merging patient profiles, health camp details, and attendance data.
- Initial exploration revealed 'Income' as a potential predictor of 'Health Score', evident in the scatter plot's trend.
- However, the predictive model, utilizing linear regression, displayed moderate R-squared values on the test set, indicating its limited ability to capture the complexities of health score prediction.
- Surprisingly, some features related to health camps or attendance seemed less impactful in predicting health scores.

Lessons Learned

- Missing data was identified, potentially influencing the model's performance, suggesting the need for imputation strategies or deeper investigation into missing value handling.
- To enhance predictive accuracy, further steps include refining feature selection, exploring additional features, and potentially employing more sophisticated modeling techniques.
- In essence, while initial insights were gained, optimizing the model by addressing missing data and refining feature selection methods is crucial to improve the accuracy of health score predictions.

THANK YOU