Healthcare Data Analysis Project: Predicting Hospital Billing Amounts Based on Admission and Patient Characteristics

1. Project Overview

1.1 Project Title:

Predicting Hospital Billing Amounts Based on Patient and Admission Characteristics

1.2 Objective:

The objective of this analysis is to predict the total billing amount for patients based on a range of variables, including patient demographics (age, gender), medical condition, insurance provider, room type, and the admission type (emergency vs. planned). Understanding the factors influencing billing amounts can help hospitals optimize cost management, improve financial forecasting, and guide resource allocation.

1.3 Scope:

- **Dataset**: The dataset contains that was sourced from kaggle.com, it does not contain real patient healthcare data.
- **Focus**: Predictive analysis to understand factors influencing billing amounts.

2. Data Collection & Preparation

2.1 Data Sources:

- **Synthetic Hospital Admission Records**: Information from the hospital's EHR(Electronic Health Record) system about patient demographics, medical conditions, doctors, admission details, billing, etc.
- **Synthetic Insurance Provider Data**: Information about the patient's insurance coverage.

- **Synthetic Billing Data**: Details of charges related to patient stay, tests, medications, and room charges.

2.2 Data Cleaning & Preprocessing:

- **Missing Data**:
  - Missing values in **Billing Amount** were removed (since it's the target variable).
- **Date Features**:
  - The **Date of Admission** and **Discharge Date** were parsed into datetime format.
- **Feature Engineering**:
  - **Age groups** were created by categorizing **Age** into bins (e.g., "18 and Under", "19-29", "30-39", "Over 80").
  - Changed the Billing Amount column data type from decimal to fixed decimal to better resemble money and offset the extra decimal spots.
  - Transformed the Name column to so that the first letter of each word would be capitalized.

2.3 Data Exploration:

- **Summary Statistics**:
  - Average **Billing Amount**: $25,500
  - **Age Distribution**: Predominantly adults 50 and up.
- **Visualizations**:
  - **A stacked area chart of Billing Amounts**: A left-skewed distribution, indicating a few high-cost treatments.
  - **Pie chart** of **Billing Amount by Insurance Provider**: Private insurance patients had a higher average billing amount compared to those with government-provided insurance.
  - **Line Graph of Billing Amount by Admission Type and Gender**: Emergency admissions had a wider range of billing amounts for female. Urgent admissions had a wider range of billing amounts for male. Elective admissions had a similar range of billing amounts for male and female.

3. Analysis Methodology

3.1 Statistical Analysis:

- **Descriptive Statistics**:
  - Calculated mean, median, and standard deviation of **Billing Amount** based on patient **Age Group**, **Gender**, and **Medical Condition**.
- **Correlation Analysis**:
  - Calculated correlations between features like **Gender** and **Billing Amount.** There may be differences in the average billing amounts for men and women, depending on treatment types

3.2 Predictive Modeling:

- **Feature Selection**:
  - Based on exploratory data analysis, important features included **Age Group**, **Medical Condition**, Gender, and **Admission Type**.
- **Training and Testing**:
  - The dataset was split into 80% training and 20% testing. Cross-validation was used to ensure that the model generalizes well.

4. Results & Findings

4.1 Key Insights:

- **Medical Condition** and **Medication** were the strongest predictors of **Billing Amount**, with **Medication** explaining most of the variation.
- Collectively **Emergency Admissions** had higher **Billing Amounts** than **Emergency Admissions**, with the average emergency admission billing amount being 4.5% higher.
- **Insurance Provider** had a little to no impact. They all were in the close to the 20% range.
- **Medication** had a noticeable impact; **Ibuprofen** patients generally had higher total billing amounts than those with **other medications**.

4.2 Statistical Significance:

- **Length of Stay and Admission Type were** statistically significant predictors of **Billing Amount**.
- **Medical Condition** showed varying levels of impact: **Diabetes, Obesity, Arthritis** were linked to higher costs.

4.3 Visualizations:

- **Feature Importance** from the model indicated that **Gender**, **Medication, Admission Type**, and **Medical Condition** were the top three factors influencing billing.
- **Line Chart plot of Billing Amount by Admission Type**: Emergency admissions showed a wider distribution of costs, with more high-billing patients.

5. Discussion

5.1 Interpretation of Results:

- The model confirmed that **elective** and **emergency admissions** result in higher costs, likely due to the need for more intensive care and more tests or procedures.
- **Medication** emerged as a key factor, suggesting that **patients prescribe Ibuprofen** may receive more expensive treatments or access to higher-cost facilities.

5.2 Limitations:

- **Insurance Coverage Data**: Some insurance provider details were missing or incomplete.
- **Medical Condition Classification**: Some conditions were generalized in the dataset, leading to possible misclassification or grouping.
- **Test Results**: This field was often incomplete, limiting its impact in the model. A lot of the data was redundant leading to close measurements.

5.3 Recommendations:

- **Cost Transparency**: Hospitals should consider providing more transparency in billing, especially for emergency admissions where the cost may vary widely.

- **Resource Allocation**: By understanding which **medical conditions** and **admission types** drive up costs, hospitals can better allocate resources to high-cost treatments.
- **Further Analysis**: Future work could involve segmenting patients based on **room type** and **medications** prescribed, which may further explain variations in **Billing Amount**. Add more fields such as ICU history, Hospital room charge, Hospital Location.

6. Conclusion

6.1 Summary of Findings:

- Key factors influencing **hospital billing** are **gender, admission type**, **medical condition**, and **insurance provider**.

6.2 Future Work:

- **Incorporating more data** on **medications** and **test results** to improve model accuracy.
- **Real-time cost prediction**: Develop a model that can predict a patient's billing amount during their stay, based off of the hospital that they are in plus their doctor.