

CostAid

Cost Management Made Easy with the Power of AI



Team Members—

- **Ali Ahmad (Team Leader)**
- **Mahima Singh**
- **Vishal**
- **Shreya Manker**
- **Soham Kamble**

Github Link - [DataEngAA/CostAid: Cost Aid is a prototype application that helps users predict and manage healthcare treatment costs \(github.com\)](https://github.com/DataEngAA/CostAid)

Problem Statement:

Hospitals and healthcare providers across different states in India frequently encounter difficulties in accurately estimating treatment costs for patients. This lack of transparency and predictability can result in financial strain for patients, billing disputes, and inefficiencies in hospital resource management. Additionally, patients often face challenges related to varying insurance coverage across states, further complicating their financial planning. Consequently, the absence of clear cost estimates and inconsistent insurance coverage impacts patients' treatment decisions and financial stability.

Market/Customer/Business Need Assessment:

This prediction system not only benefits the patient, it comes out to be beneficial for all the members which are part of this problem statement apart from patients. By considering following points we can conclude how it could be beneficial for Business, Market & customer perspective.

- **Cost Transparency:** Patients often face uncertainty about the costs of medical treatments, leading to financial stress and dissatisfaction.
- **Cost Management:** Hospitals struggle with accurately estimating treatment costs, leading to budget overruns and inefficient resource allocation.
- **Claims Processing:** Insurance companies deal with inconsistencies in treatment cost estimations, resulting in delayed and disputed claims.
- **Market Competition:** Healthcare providers need to stay competitive by offering transparent and predictable pricing.

Target Segments and Market needs:

1. Hospitals and Healthcare Providers

Healthcare Providers can leverage this model to:

- Provide transparent pricing to patients, helping build trust.
- Aid in pre-treatment financial planning, allowing patients to make informed decisions.
- Optimize bed allocation and treatment planning, improving resource management.

- Use the tool for cost benchmarking against industry averages to identify areas for cost reduction and operational improvement.

2. Insurance Companies

Insurance providers can use this model to:

- Predict claims expenses accurately and adjust premiums accordingly.
- Incorporate cost estimations into risk assessments for more tailored insurance policies.
- Enhance customer satisfaction and retention by preventing underpricing and overpricing.
- Employ the model's predictive capabilities to detect anomalies in claims, potentially identifying fraudulent activities.

3. Pharmaceutical Companies and Medical Suppliers

Pharmaceutical companies can utilize these services to:

- Anticipate demand for medications and supplies based on treatment plans.
- Adjust production and supply chain strategies proactively.
- Set competitive pricing strategies for products by understanding cost dynamics in treatment processes.

4. Healthcare Policy Makers and Regulators

Policy makers can use the model as a tool for:

- Monitoring and managing healthcare costs at a macro level.
- Analyse data trends and cost implications of various treatment protocols for informed decision-making.
- Ensure equitable access to healthcare services and craft policies that encourage cost-effective treatment practices.

5. Common Customers (Patients and Caregivers)

Common customers can:

- Gain a clearer understanding of potential medical expenses before treatment.
- Make better-informed decisions about healthcare options, budgeting, and insurance needs.
- Alleviate financial uncertainty and focus on recovery rather than worrying about unexpected medical bills.

6. Research Institutions and Academic Bodies

- Leverage data and insights from the model to study healthcare economics and disease patterns.
- Inform studies on the economic impact of different healthcare strategies.
- Contribute to the development of more efficient and cost-effective healthcare solutions.

7. Telemedicine and Remote Healthcare Providers

- Use the model as a tool for offering comprehensive care, integrating cost predictions into service offerings.
- Enhance value proposition by providing financial guidance alongside medical advice.

Cater to patients seeking remote consultations with a detailed cost analysis.

8. Medical Tourism Agencies

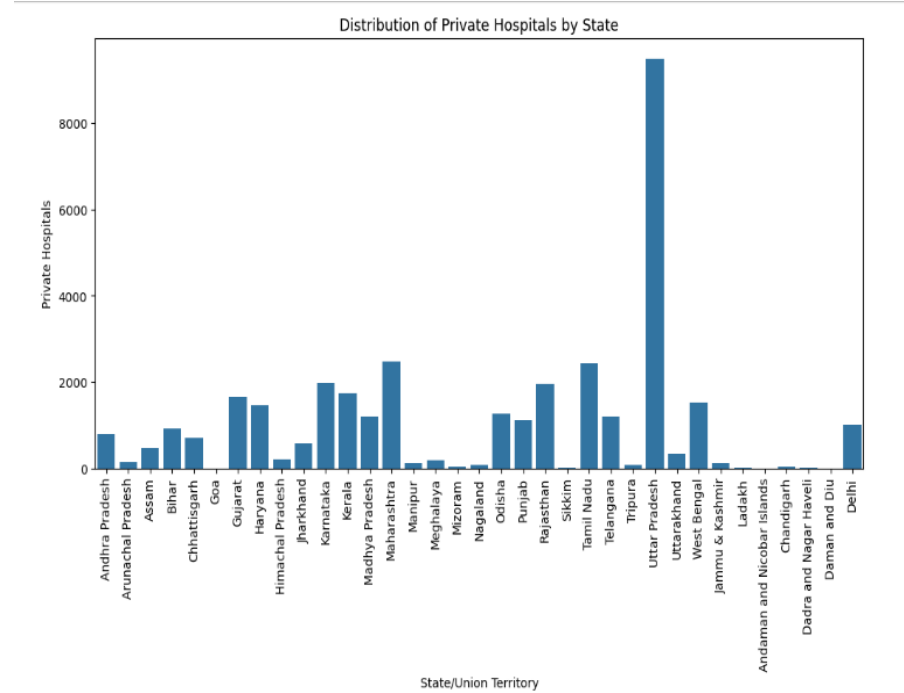
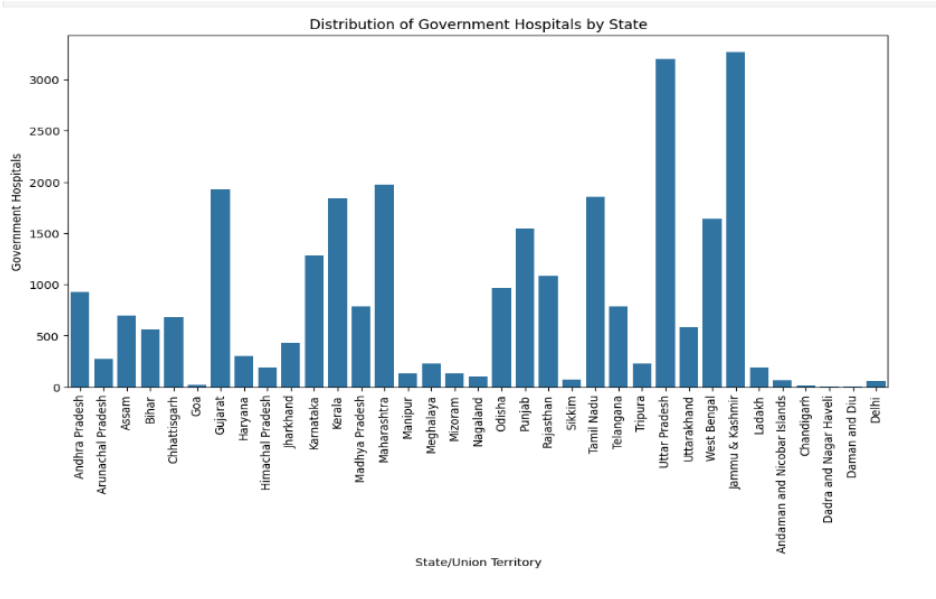
- Offer potential clients a detailed cost analysis of treatments in different locations.
- Enhance service offerings by providing transparent pricing.
- Build trust with clients and assist in planning travel logistics and budgeting.

▪ **Hospital Data Analysis:**

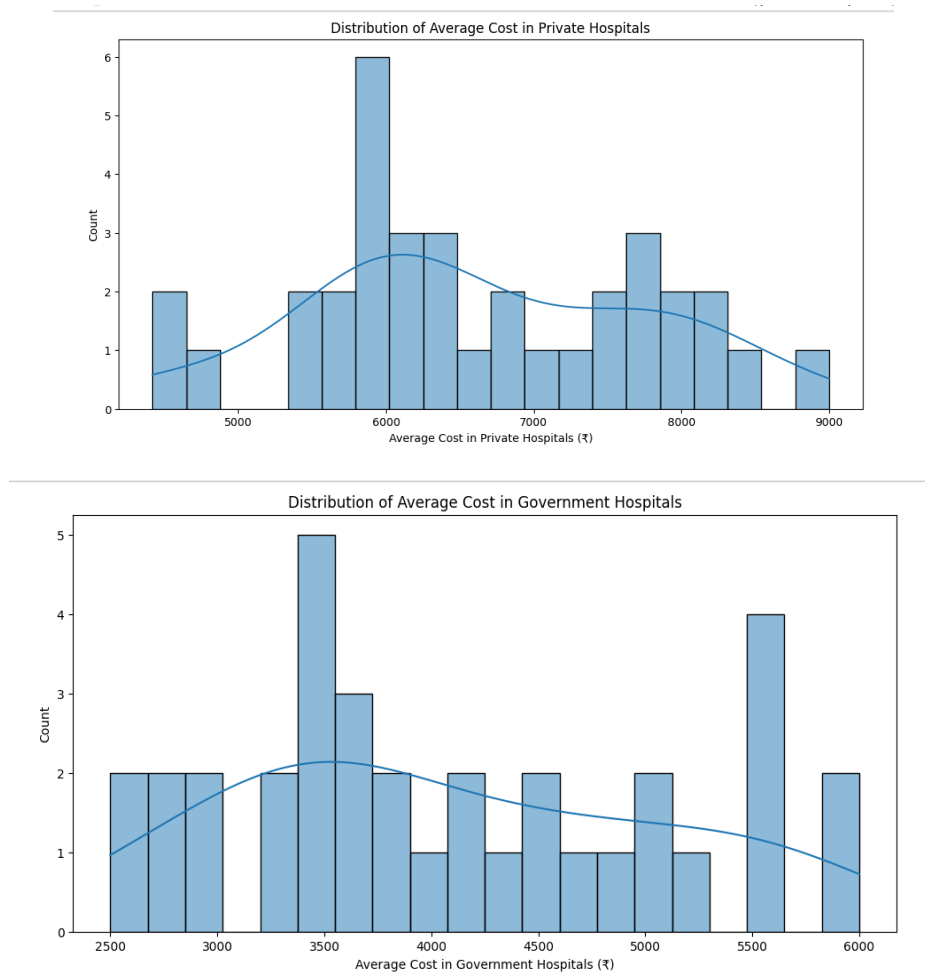
An analysis of hospital state data, including the number of private and government hospitals, average costs in these hospitals, patient preferences for homeopathy versus conventional medicine, and age group distribution. The goal is to understand the current landscape of healthcare facilities, cost variations, patient preferences, and demographic trends in healthcare utilization.

Hospital Distribution and Costs-

The analysis covers the number of hospitals (private and government) across different states and the average cost of treatment in these hospitals.

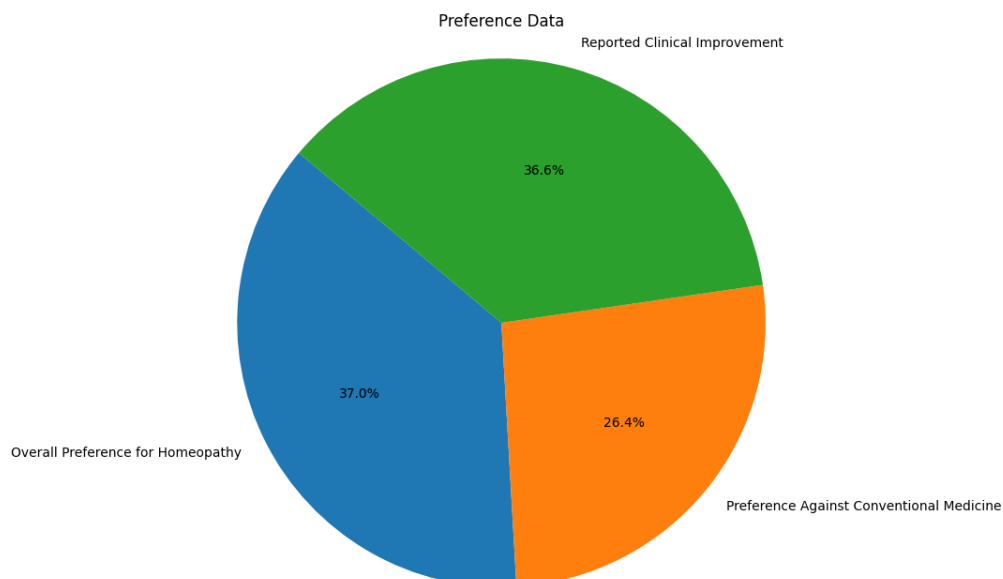


Average Costs in Hospitals (State-wise)-



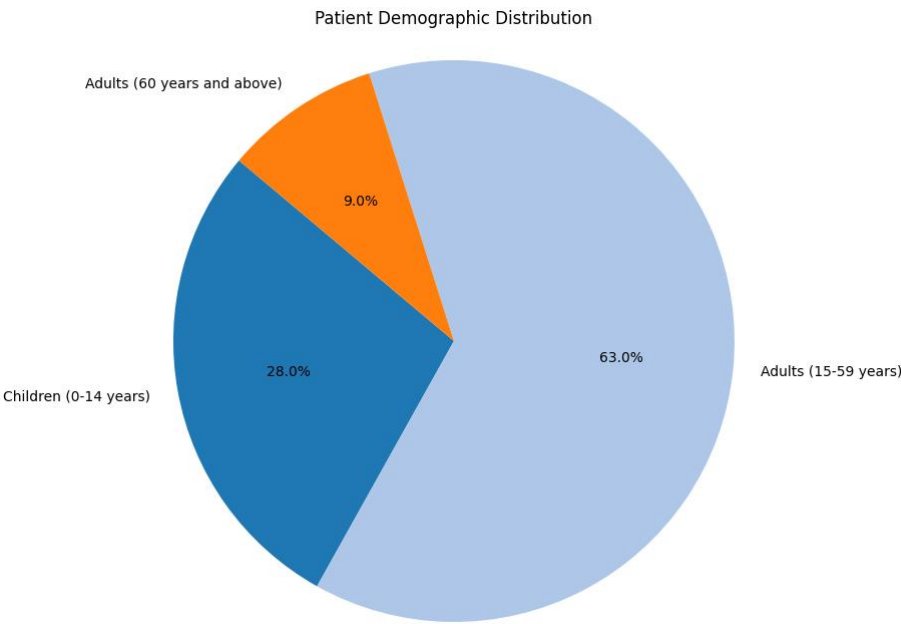
Patient Preferences and Homeopathy-

Homeopathy has gained significant preference among patients for various reasons, including a holistic approach, fewer side effects, and dissatisfaction with conventional medicine.



Age Group Analysis-

Understanding which age groups are most frequently seeking medical care helps tailor healthcare services and resources effectively.



- Cost Distribution by Age Group-

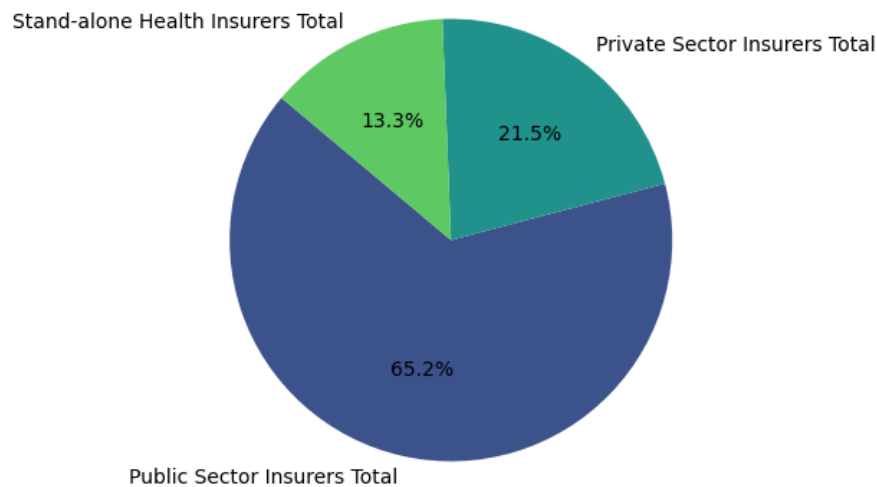
Age Group	Average Cost in Private Hospitals (₹)	Average Cost in Government Hospitals (₹)
Children (0-14 Years)	5,000 - 8,000	2,500 - 4,500
Young Adults (15-24 Years)	6,500 - 9,000	2,800 - 5,000
Adults (25-64 Years)	8,000 - 12,000	3,000 - 5,500
Seniors (65 Years and Above)	10,000 - 15,000	3,500 - 6,000

The data highlights significant variations in healthcare costs and facilities across states, a strong preference for homeopathy among certain demographics.

This report provides a comprehensive view of the current healthcare landscape, enabling stakeholders to make informed decisions to improve healthcare delivery and accessibility in India.

▪ Evaluation of Customer Profile for Insurers:

• Number of People Covered in 2021-22 by Insurer Category:



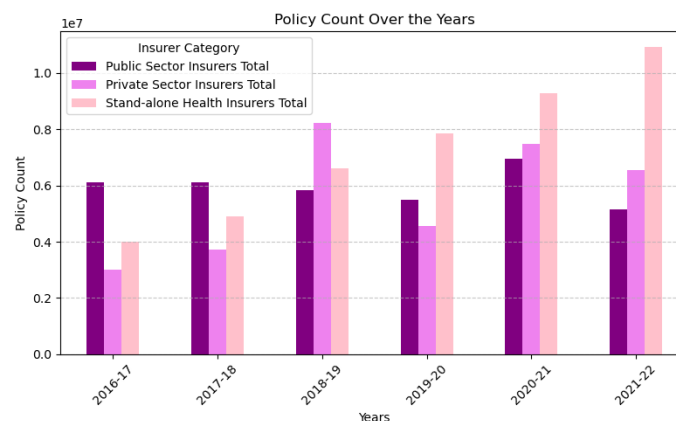
• Market Share Analysis:

- Public Sector Insurers cover the largest portion of people (65.2%), followed by Private Sector Insurers (21.5%) and Stand-alone Health Insurers (13.3%).

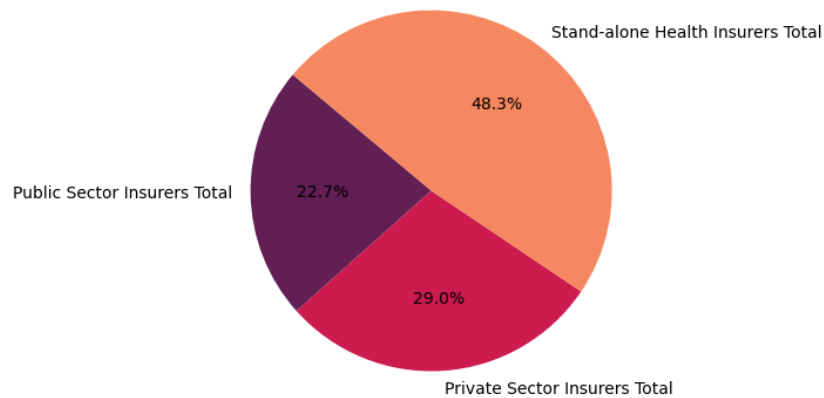
• Implication:

- The high market share of Public Sector Insurers means they have a large, stable customer base, making them a key target for the AI model.
- Private Sector Insurers, with a notable share, are potential customers needing sophisticated tools to remain competitive and optimize costs.
- Stand-alone Health Insurers, though covering fewer people, are rapidly growing, indicating potential future demand.

Policy Count over the years by Insurer Category:



Policy Count in 2021-22 by Insurer Groups



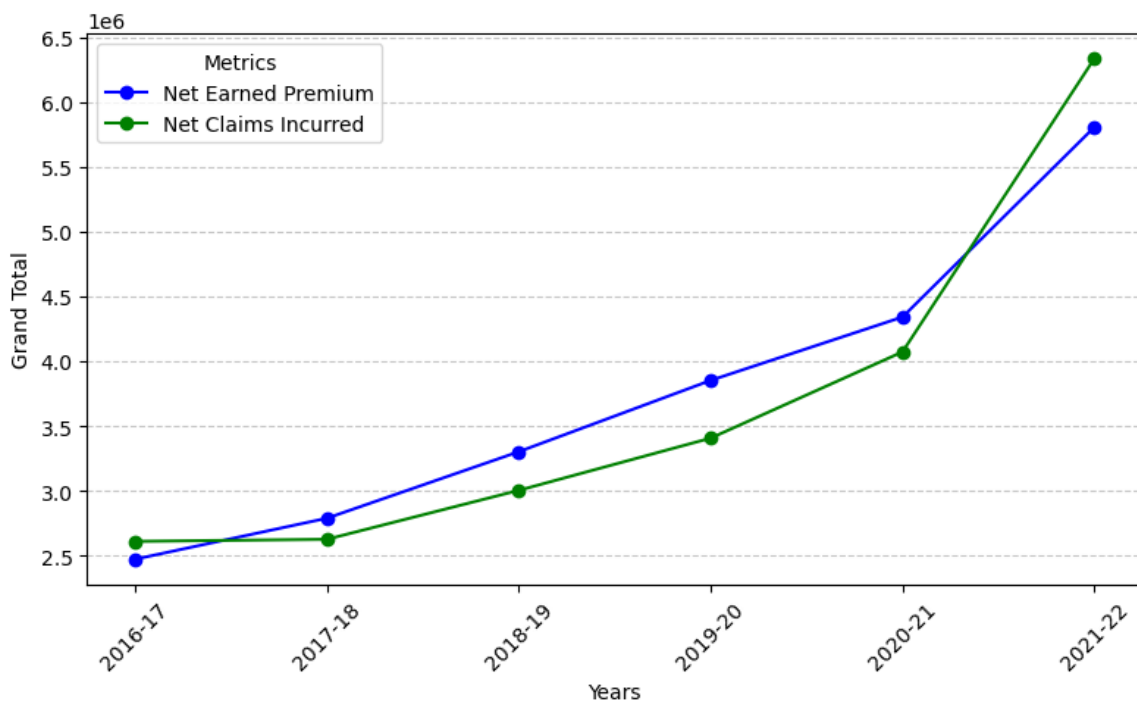
- **Policy Distribution:**

- Stand-alone Health Insurers issued the most policies (48.3%) in 2021-22, followed by Private Sector Insurers (29.0%) and Public Sector Insurers (22.7%).

- **Implication:**

- Stand-alone Health Insurers, despite covering fewer people, consistently issue more policies over the years, indicating a preference for smaller, targeted customer segments. This suggests a need for more granular and precise cost predictions, making them prime candidates for advanced features of the AI model.

Net Earned Premium vs. Claims Incurred Over the Years:

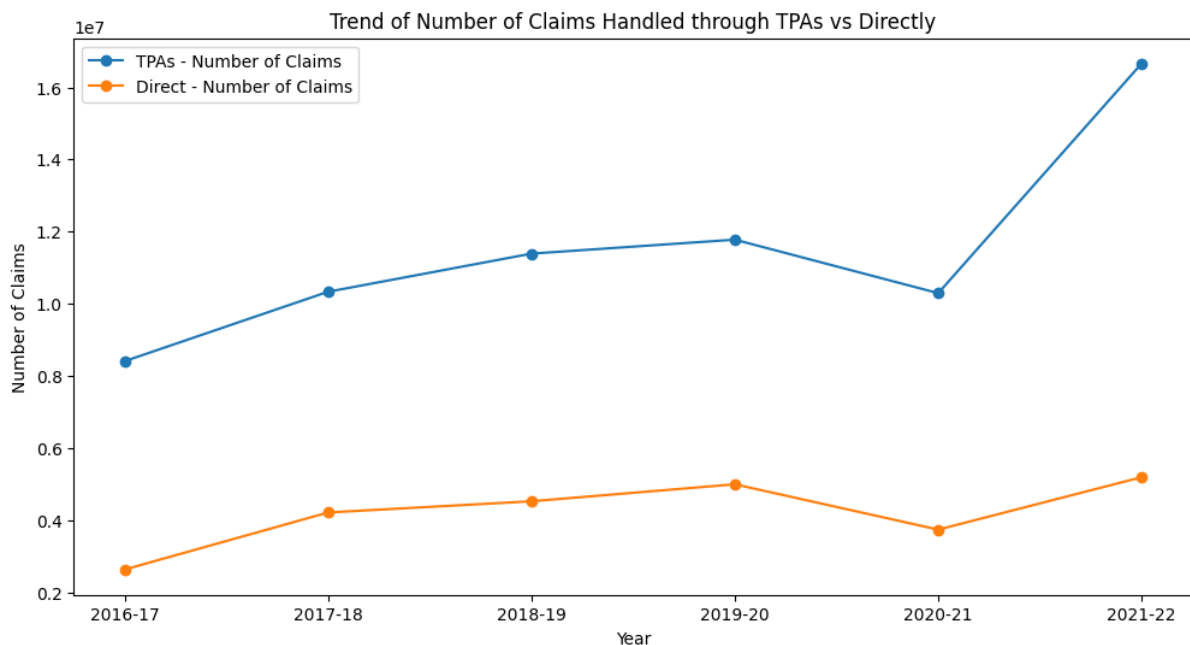


- **Financial Performance:**

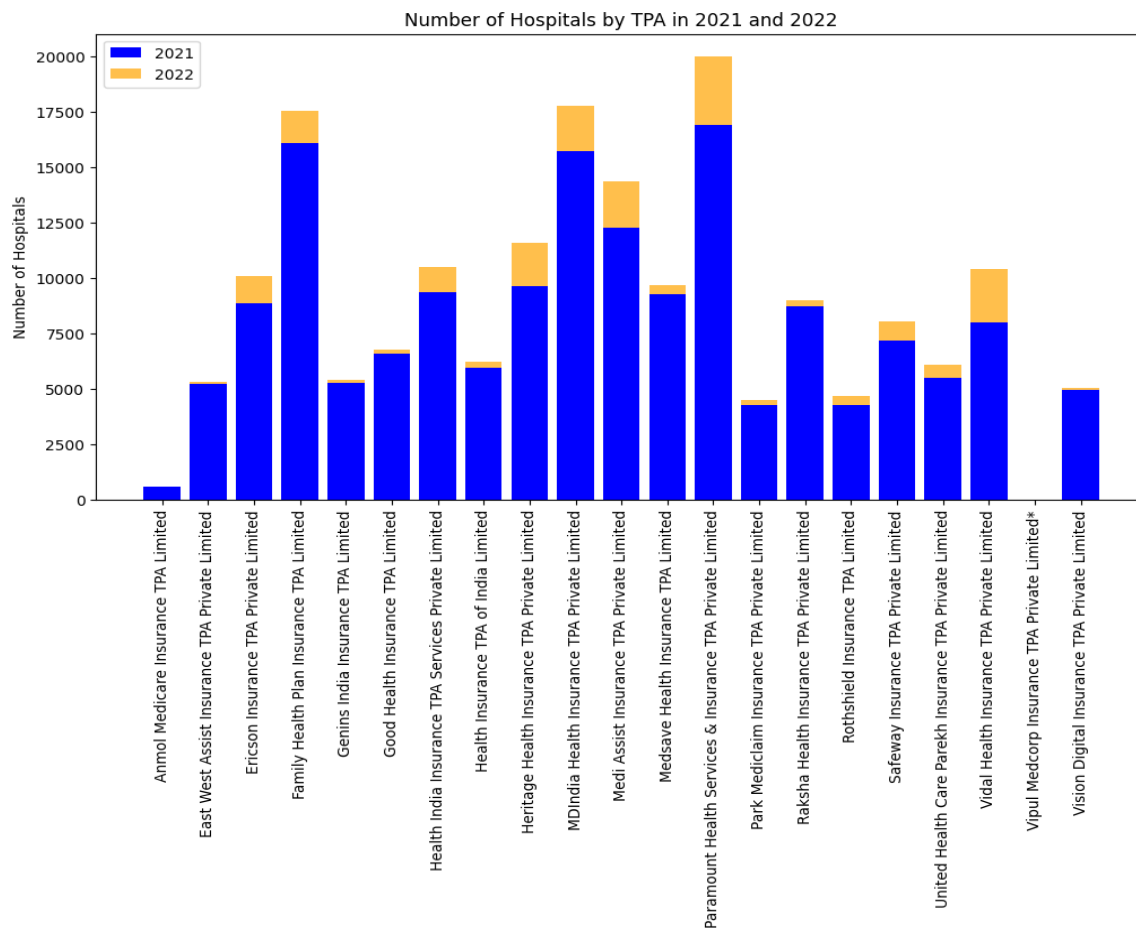
- Both Net Earned Premiums and Net Claims Incurred have risen consistently over the years, with a sharp increase in 2021-22.
- In 2021-22, Net Claims Incurred outpaced Net Earned Premiums, indicating higher risk and potential financial strain for insurers.

- **Implication:**

- Insurers, particularly those seeing a rise in claims, would benefit from predictive tools that can help manage costs better. The AI model can provide cost optimization by predicting treatment expenses, allowing insurers to better align premiums with potential claims.
- There is an opportunity to offer detailed analyses and risk management features to insurers experiencing higher claims, particularly targeting those needing to balance premiums and payouts.



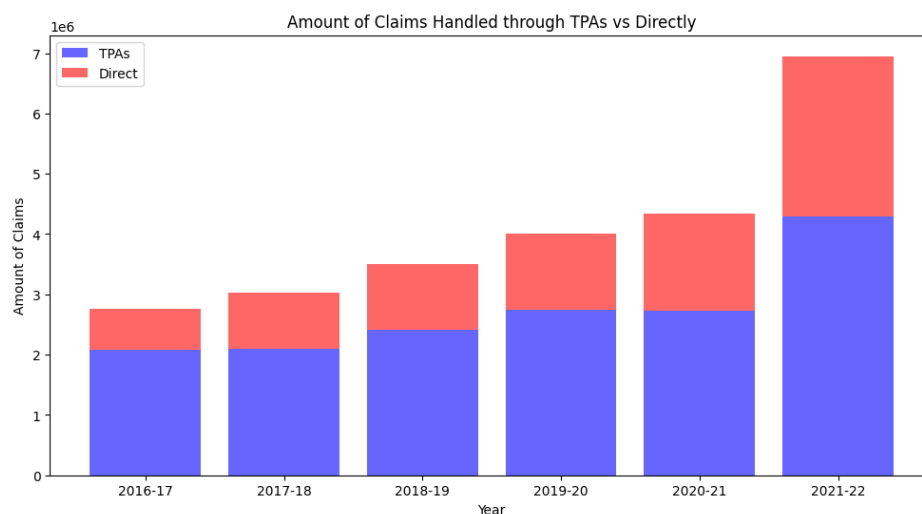
- Insurers are increasingly relying on TPAs to manage claims. This suggests that TPAs are a critical intermediary for insurers, making them a key customer segment.



There is a visible increase in the number of hospitals managed by most TPAs from 2021 to 2022, indicating an expansion of healthcare networks.

□ This growth suggests an increasing demand for services like predictive cost modelling. The expanding TPA networks can significantly benefit from the detailed insights and cost predictions your AI service provides, as they manage more healthcare institutions over time.

▪ **Steadily Rising Claims Amount:**



- The increasing claim amounts highlight the growing financial burden on insurers. This trend makes cost prediction even more critical. Insurers and TPAs would be interested in your model to anticipate and manage these rising costs more effectively.

Overall Customer Profile Insights:

- **Public Sector Insurers:** Major players with vast coverage and significant policy counts. They need broad, reliable predictive tools.
- **Private Sector Insurers:** Strong middle-ground players with a substantial market share, requiring advanced analytics to stay competitive.
- **Stand-alone Health Insurers:** Fast-growing with a high number of policies issued, indicating a need for flexible, specialized tools for smaller customer segments.
- **TPAs:** High-volume users that require robust, scalable solutions for managing large numbers of claims. They are likely to benefit from enterprise-level subscription plans.
- **Direct Insurers:** Smaller but stable market segment that may need more customized or mid-level solutions. They could benefit from professional or business-tier subscription plans.

Bench Marking Alternate Products

Product benchmarking is an important step as it analyse your product's performance and functionality against competitors to identify areas of improvement and enable strategic development of your product roadmap. It also helps you define key differentiators and establish your competitive positioning. To benchmark the Hospital Treatment Pricing Prediction System (HTPPS), it's important to identify and analyse existing products and services in the market that offer similar capabilities.

Clear Health Costs •

Description: Clear Health Costs is a platform that provides price transparency for healthcare services by collecting and sharing crowdsourced pricing data.

• Features:

Crowdsourced data collection from patients.

Searchable database of treatment costs.

Price comparisons for various medical procedures.

IBM Watson Health Payer Analytics

- **Description:**

IBM Watson Health offers a suite of analytics tools designed for payers, including cost prediction models.

- **Features:**

Predictive analytics for cost management.

Integration with existing healthcare IT systems.

Advanced machine learning algorithms.

Business Model

Free Plan: Basic Access

- **Description:** The free plan allows new users to test the service and see its benefits.
- **Features:**
 - 3 free basic cost predictions per month
 - Access to a limited number of disease and symptom analyses
 - Customer support via email
- **Target Users:** Casual users, patients trying to understand potential costs, and first-time visitors

Subscription Plan 1: Starter

- **Price:**
 - Monthly: ₹1,499
 - Annual: ₹14,999 (saves ~16%)
- **Features:**
 - 10 cost predictions per month
 - Basic disease and symptom analyses
 - Option to save and review previous analyses
- **Target Users:** Patients and caregivers who need more than occasional use

Subscription Plan 2: Professional

- Price:
 - Monthly: ₹3,999
 - Annual: ₹39,999 (saves ~16%)
- Features:
 - 30 cost predictions per month
 - Detailed disease and symptom analyses with recommended treatment plans
 - Cost breakdown by treatment categories (e.g., medications, bed charges)
 - Ability to share analyses with healthcare providers
- Target Users: Insurance agents, healthcare professionals, and patients with ongoing needs

Subscription Plan 3: Business

- Price:
 - Monthly: ₹9,999
 - Annual: ₹99,999 (saves ~16%)
- Features:
 - 100 cost predictions per month
 - Advanced disease and symptom analyses with real-time updates
 - Customizable reports and analytics for internal use
 - Integration with existing Customer Relationship Management (CRM) and healthcare management systems
- Target Users: Small to medium-sized hospitals, clinics, and medical practices

Subscription Plan 4: Enterprise

- Price:
 - Monthly: ₹24,999
 - Annual: ₹2,49,999 (saves ~16%)
- Features:

- 300 cost predictions per month
- Comprehensive disease and symptom analyses including predictive insights and forecasting
- Customized dashboards and KPIs (Key Performance Indicators)
- API access for seamless integration with internal systems
- Target Users: Large hospitals, healthcare systems, and insurance companies

Subscription Plan 5: Unlimited

- Price:
 - Annual: ₹4,99,999
- Features:
 - Unlimited cost predictions
 - Full access to all features, analyses, and reports
 - Regularly scheduled performance reviews and optimization sessions
- Target Users: Major healthcare organizations, large insurance corporations, and global medical networks

Model Generalization

Dataset Description

The dataset which we have used in our model building comprises of almost 4.9 lakhs of record, it is based upon a medical survey which has been conducted in USA.

It comprises of total 17 feature variables and 1 target variable as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 490000 entries, 0 to 489999
Data columns (total 18 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Age Group                                490000 non-null  object
1   Gender                                  490000 non-null  object
2   Length of Stay                          490000 non-null  int64
3   Type of Admission                       490000 non-null  object
4   Patient Disposition                     490000 non-null  object
5   CCS Diagnosis Description                490000 non-null  object
6   CCS Procedure Description               490000 non-null  object
7   APR DRG Description                     490000 non-null  object
8   APR MDC Description                     490000 non-null  object
9   APR Severity of Illness Description      490000 non-null  object
10  APR Risk of Mortality                   490000 non-null  object
11  APR Medical Surgical Description         490000 non-null  object
12  Payment Typology 1                      490000 non-null  object
13  Payment Typology 2                      315687 non-null  object
14  Payment Typology 3                      136615 non-null  object
15  Birth Weight                            490000 non-null  int64
16  Emergency Department Indicator           490000 non-null  object
17  Cost INR                                490000 non-null  float64
dtypes: float64(1), int64(2), object(15)
memory usage: 67.3+ MB
```

Generalized Model

After implementing various regression models, we have considered '*Random Forest Regression*' as the best fitted model for our dataset.

In the initial step we have split the data in Training and Testing set, then we have performed fitting and prediction over the regression model as follows:

- I. Creation of regression model after performing hyper parameter tuning and

training the regressor over the training set.

5. Random Forest Regressor

```
[61]: rf_model = RandomForestRegressor(n_estimators=100,max_depth=6, random_state=42)
      rf_model.fit(X_train, y_train)
```

```
[61]: RandomForestRegressor(max_depth=6, random_state=42)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.

On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

☒ [RandomForestRegressor?Documentation for RandomForestRegressor](#)
Fitted

```
RandomForestRegressor(max_depth=6, random_state=42)
```

- II. Performing prediction using the test dataset and evaluation of the test score.

```
[62]: y_pred_rf=rf_model.predict(X_test)
      r2_score(y_test,y_pred_rf)
```

```
[62]: 0.6327492122615432
```

- III. Checking the over-fitting of the regression model.

>>> Check for Over-fitting

```
[64]: cv_scores_rf = cross_validate(rf_model, X_train, y_train, cv=10, scoring="r2",return_train_score=True)
```

```
[65]: print("Test>>>", "Min:",cv_scores_rf.get("test_score").min(), "Max:",cv_scores_rf.get("test_score").max())
      print("="*60)
      print("Train>>>", "Min:",cv_scores_rf.get("train_score").min(), "Max:",cv_scores_rf.get("train_score").max())
```

```
Test>>> Min: 0.5985728313692291 Max: 0.6843897369638008
```

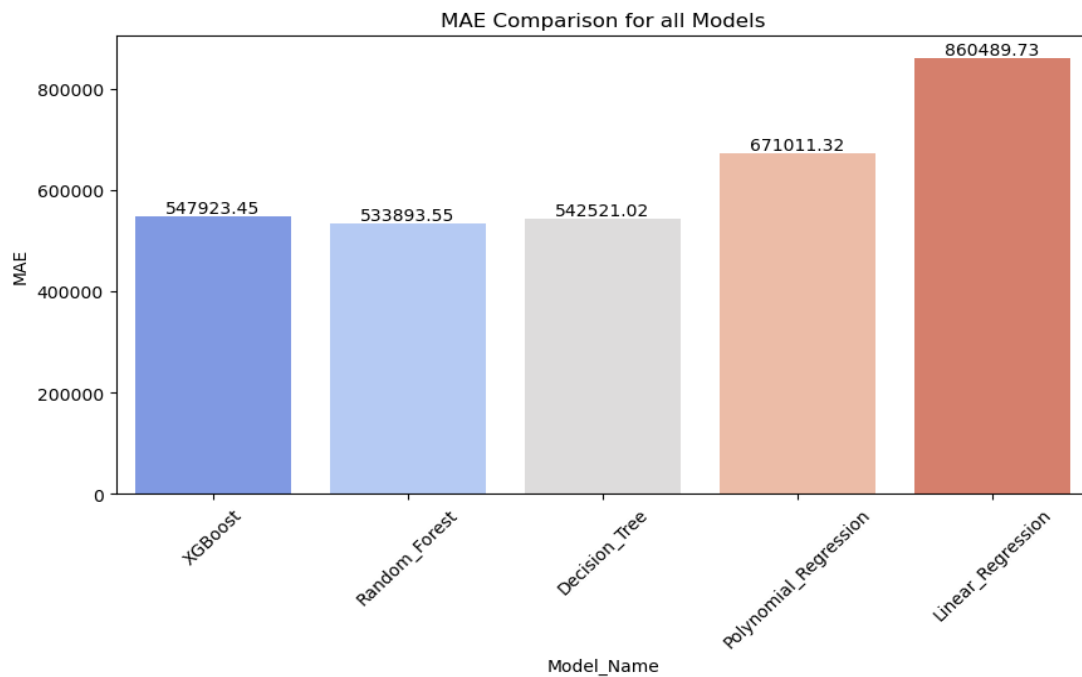
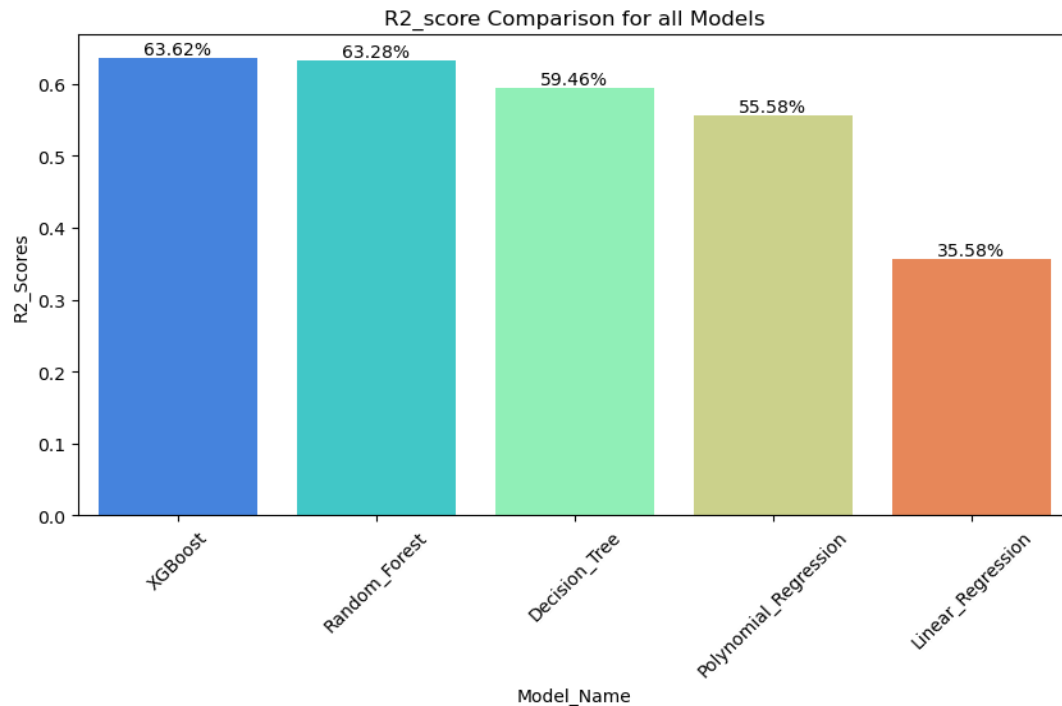
```
=====
```

```
Train>>> Min: 0.7102635832993396 Max: 0.7148813780915088
```


Why this model?

We have performed predictions using various regression model such as XGBoost, Random Forest, Decision Tree, Polynomial Regression and Linear Regression.

Among them all Random Forest comes out to be the best as compare to other models it has the best prediction *accuracy score*, *least mean absolute error* and the issue of getting *overfit* is *slightly as less* compared other regression model.

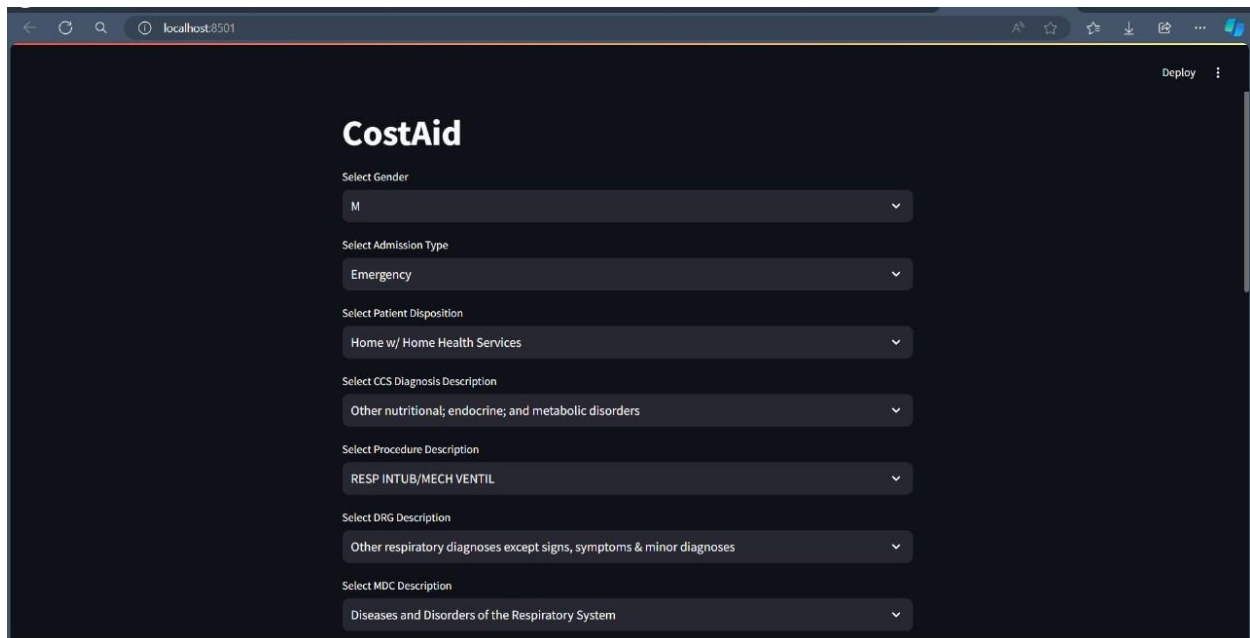


Conclusion

After considering all the regression models and comparing their accuracy score, mean absolute error, Prediction Vs actual plots and residual plots we have come to the conclusion of using **Random Forest as our final Prediction Model**.

After going through multiple iterations of Hyper-parameter tuning, we have reached to this accuracy score for **Random Forest Regressor (around 63%)**.

Application Prototype:



The screenshot shows a web browser at localhost:8501 displaying the 'CostAid' application. The interface has a dark theme. The title 'CostAid' is at the top left. Below it, there are eight dropdown menus for selecting various medical and administrative details. The 'Deploy' button is in the top right corner.

CostAid

Select Gender: M

Select Admission Type: Emergency

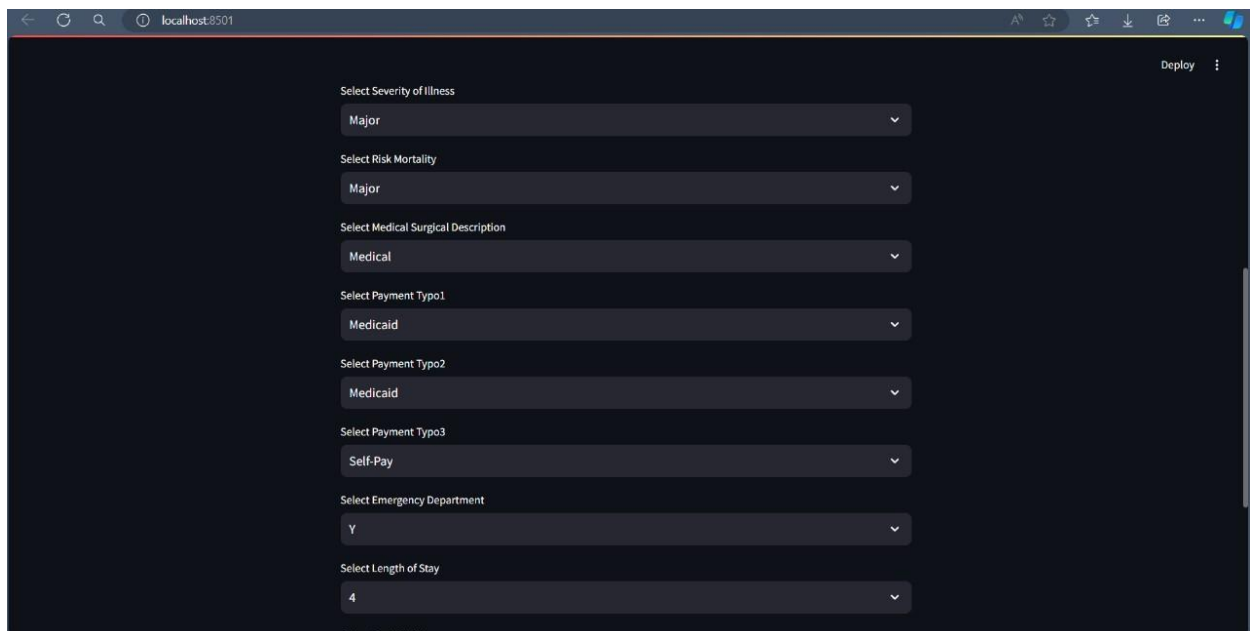
Select Patient Disposition: Home w/ Home Health Services

Select CCS Diagnosis Description: Other nutritional, endocrine, and metabolic disorders

Select Procedure Description: RESP INTUB/MECH VENTIL

Select DRG Description: Other respiratory diagnoses except signs, symptoms & minor diagnoses

Select MDC Description: Diseases and Disorders of the Respiratory System



The screenshot shows the same web browser at localhost:8501, but with a different set of dropdown menus. The 'Deploy' button remains in the top right corner.

Select Severity of Illness: Major

Select Risk Mortality: Major

Select Medical Surgical Description: Medical

Select Payment Typo1: Medicaid

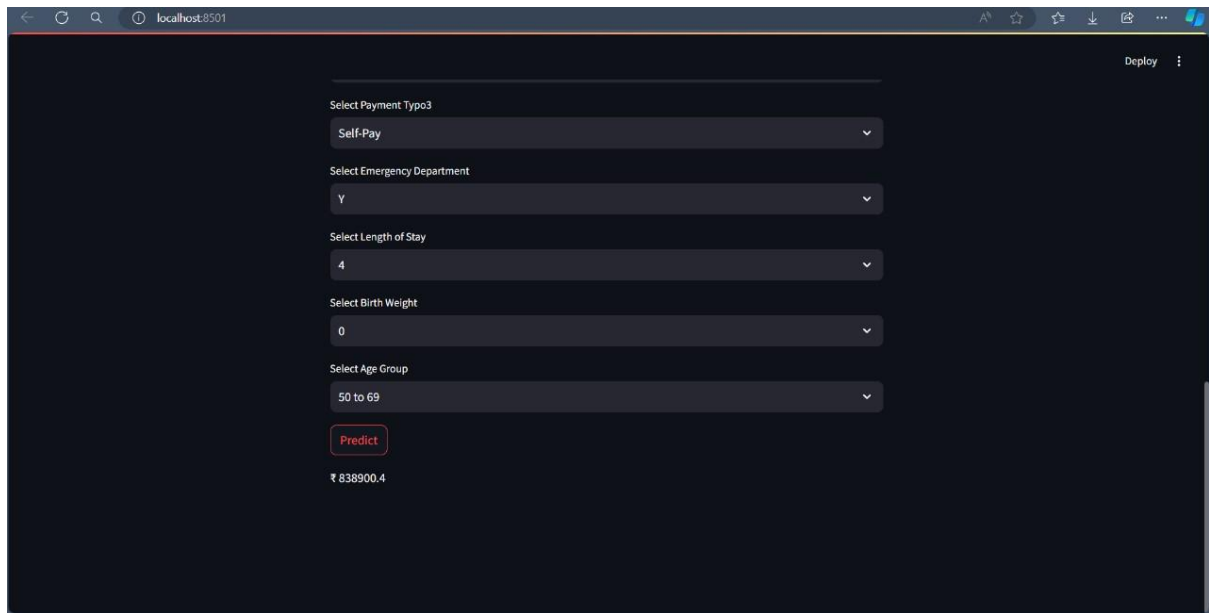
Select Payment Typo2: Medicaid

Select Payment Typo3: Self-Pay

Select Emergency Department: Y

Select Length of Stay: 4

Calculate Risk Weight



Basic Financial Equation:

To create a financial equation that represents the annual profits for our AI predictive model, we can use the format

$$\pi_t = m * x_t - c_t$$

Where:

π_t = total profit in year 't'

m = price of the services

x_t = total sales as in year 't'

c_t = total cost in year 't'

We can further modify this equation as:

$$\pi_t = R_t - C_t$$

Where:

R_t = total revenue in year 't'

C_t = total cost in year 't'

Variable:

N_{SP1} : Number of users subscribed to Starter Plan for year 't'

N_{SP2} : Number of users subscribed to Professional Plan for year 't'

N_{SP3} : Number of users subscribed to Business Plan for year 't'

N_{SP4} : Number of users subscribed to Enterprise Plan for year 't'

N_{SP5} : Number of users subscribed to Unlimited Plan for year 't'

P_{SP1} : Annual price of Starter Plan

P_{SP2} : Annual price of Professional Plan

P_{SP3} : Annual price of Business Plan

P_{SP4} : Annual price of Enterprise Plan

P_{SP5} : Annual price of Unlimited Plan

C_{fixed} : Annual fixed costs (e.g., infrastructure, salaries, marketing)

$C_{variable}$: Annual variable costs (e.g., customer support, data storage) per user

Revenue Calculation:

The total revenue R for year 't' from all subscription plans is calculated as:

$$R_t = (N_{SP1} * P_{SP1}) + (N_{SP2} * P_{SP2}) + (N_{SP3} * P_{SP3}) + (N_{SP4} * P_{SP4}) + (N_{SP5} * P_{SP5})$$

Cost Calculation:

The total costs C include fixed costs and variable costs:

$$C_t = C_{fixed} + (N_{SP1} + N_{SP2} + N_{SP3} + N_{SP4} + N_{SP5}) * C_{variable}$$

Final Financial Equation:

Profit in year 't' is the difference between the total revenue and costs incurred in that year.

$$\pi_t = (N_{SP1} * P_{SP1}) + (N_{SP2} * P_{SP2}) + (N_{SP3} * P_{SP3}) + (N_{SP4} * P_{SP4}) + (N_{SP5} * P_{SP5}) - \{C_{fixed} + (N_{SP1} + N_{SP2} + N_{SP3} + N_{SP4} + N_{SP5}) * C_{variable}\}$$

The equation emphasizes the need to balance competitive pricing with premium features that justify higher subscription rates, particularly for higher-tier plans. Furthermore, it highlights the importance of controlling variable costs to ensure they do not erode profitability as the user base grows. With careful management, the derived equation serves as a robust framework for forecasting and optimizing financial performance in a rapidly expanding market.

Market Growth Forecast:

Over the next five years, we anticipate robust growth for our AI hospital price predictive model, driven by the increasing demand for cost transparency and efficiency in the healthcare sector. The rise in healthcare costs, coupled with the expanding adoption of AI and data analytics, will significantly boost the need for predictive tools among insurers, hospitals, and healthcare providers. As health insurance penetration grows and the digital health ecosystem expands, our product will become essential for managing costs, optimizing premiums, and enhancing patient care. Additionally, supportive government initiatives and favourable regulations for AI in healthcare will further accelerate market adoption, positioning our model as a critical tool in the industry.

Considering these factors, we project a compound annual growth rate (CAGR) of 15-20% for our product over the next five years. This growth will be driven by increased subscriptions from diverse segments. By capitalizing on these market dynamics, we expect our AI model to see widespread adoption, leading to substantial revenue growth and a strong market presence in the healthcare sector.

Conclusion

Our analysis of the healthcare and insurance landscape across India reveals significant insights into the treatment cost dynamics and insurance coverage patterns. Here are the key findings:

1. Hospital Distribution and Costs:

- There is a notable split between private and government hospitals, with 55.08% of hospitals being private and 44.02% government-run.
- Costs for treatments vary significantly between these sectors. Private hospitals generally have higher costs compared to government hospitals, with considerable variability based on age groups.

2. Insurance Coverage and Expenditure:

- Health insurance coverage stands at 41% of households, with 58.7% of total health expenditure being out-of-pocket.
 - The public sector holds the largest market share in insurance, while stand-alone health insurers are growing rapidly.
3. Demographics and Preferences:
- The majority of patients are adults aged 15-59 years, with a high preference for homeopathy over conventional medicine.
 - The cost of treatment varies widely across states, with some states like Arunachal Pradesh and Jharkhand showing lower government hospital costs.
4. Market Share and Policy Distribution:
- Public Sector Insurers cover the largest portion of the population, followed by Private Sector and Stand-alone Health Insurers.
 - Stand-alone Health Insurers issue the most policies, suggesting a preference for targeted customer segments and indicating a need for precise cost predictions.
5. Financial and Network Expansion:
- There has been a rise in Net Claims Incurred surpassing Net Earned Premiums, reflecting a higher risk for insurers.
 - The increase in the number of hospitals managed by TPAs from 2021 to 2022 underscores a growing need for advanced predictive tools.
6. Model Selection and Performance:
- After evaluating various regression models, **Random Forest** was selected as the final prediction model due to its accuracy and performance metrics. The model achieved an accuracy score of approximately 63% after hyper-parameter tuning.

The insights gained from this analysis highlight the need for advanced predictive tools to manage treatment costs and insurance claims effectively. The expanding healthcare networks and varying state-level costs present significant opportunities for leveraging AI in optimizing healthcare resource management and financial planning.