Real-World Predictive Analytics Challenge (RWPAC)

Team Name: Team Two (AKA the Mountain Hawk Med)
Team Members: Milonee Parekh, Wei Fang, Josh Garcia
Shared Google Drive Folder Link: here
Phase 4 Folder Link: here
Canva Presentation: here

Defining the Problem:

Healthcare organizations struggle to forecast hospitalization costs due to the complexity of patient profiles, comorbidities, and variations in treatment requirements. This lack of accurate cost predictions can lead to budget overruns, inefficient resource utilization, and even under- or over-allocation of medical staff and supplies. This can also be stressful for the patient because they do not have an accurate estimate of how much they will have to pay. As a result, organizations face challenges in financial planning, effective resource management, and maintaining patient care standards.

Moreover, a 2019 study by the **American Cancer Society** found that *56% of adults reported experiencing at least one cost-related problem* with their healthcare or health insurance in the past year, such as delaying or skipping care due to costs. This highlights a critical issue: when healthcare costs are unpredictable or poorly managed, it has a direct impact on patient health and access to care. Solving this problem would not only benefit healthcare organizations but also improve the affordability and accessibility of care for patients.

Dataset:

We were able to collect our data from Kaggle. There are three tables that can be joined into a single data frame based on the ‘Customer ID’ value.

Hospitalization Details Data Set: link

Medical Examinations Data Set: link

Names Data Set: link

Upon initial inspection of the combined data set, the data is usable and reliable. The main data types are integers and strings. From observation, we noticed that maybe we can split the name column into multiple columns including prefix, first name, and last name. Also, the word “tier” could potentially be removed from the “Hospital tier” and “City tier” columns. It is redundant, it will make it difficult when we are trying to perform calculations. An alternative is we could turn

them into dummy variables. This would solve the problem and we would get practice converting a column into multiple columns for dummy variables.

Initial Analysis

Continuous Variables (e.g.charges, BMI, HBA1C):

Charges:

- The average (mean) hospital charge is **13,529.92**.
- The median is **9,630.91**, indicating the central tendency.
- The range of charges is from **563.84** to **63,770.43**, indicating a significant spread in the charges across patients.

BMI:

- The average Body Mass Index (BMI) is **30.97**, suggesting many patients fall in the overweight or obese category.
- The BMI range is from **15.01** to **55.05**, indicating extreme variation from low to high BMI values.

HBA1C:

- The average HBA1C (a measure of blood sugar) is **6.58**.
- The HBA1C range is from **4.00** to **12.00**, reflecting some elevated levels indicating diabetes cases.

Age:

- The average age is 40
- The ages range from 19 to 66

Categorical Variables (e.g., Hospital tier, City tier, Heart Issues, etc.):

Hospital Tier:

- Tier 2 hospitals are the most common, followed by Tier 3. Tier 1 hospitals are less frequent.

City Tier:

- Most patients come from Tier 1 and Tier 3 cities, suggesting a wide geographical distribution.

Heart Issues:

- A majority of patients (**1409**) report no heart issues, though a smaller portion (**926**) have heart issues.

Any Transplants:

- Most patients (**2191**) have not had transplants, with only a small percentage (**144**) reporting a history of transplants.

Cancer History:

- A large majority (**1944**) have no history of cancer, though a smaller group (**391**) does.

Smoker:

- A significant portion of patients are non-smokers (**1845**), though a smaller but notable proportion (**488**) are smokers.

Data Cleaning Tasks:

(Explain decisions made during data cleaning)

Mountain Hawk Med understands that data cleaning is an essential part of data analytics. It ensures that the data is accurate, consistent, and high quality. Cleaning data allows the analysis portion of the project to run smoother because using clean data can help remove errors that would have come up in the future, in turn, saving time and resources. Also, accurate data leads to more reliable conclusions and insights.

Number Of Major Surgeries:

With that being said, Mountain Hawk Med cleaned our data a few different ways. The first thing that we did was clean the “NumberOfMajorSurgeries” column. The values for this column were “No major surgery”, “1”, “2”, and “3”. We decided that it would be more efficient to change all of the “No major surgery” values to “0” so that it better aligns with the format of the other answers.

Replacing NaN Values:

We also cleaned all of the “?” values in the data set. Whenever a value was uncertain, the “?” character was used as a placeholder indicating that the value is undetermined. This is similar to the NaN value. Instead of just changing those values to NaN and ignoring those rows, we decided to replace those “?” values with the mode of that column. This would allow us to still be able to use all of the data in the set without having to ignore any rows because they have NaN values. Making this change allowed our visualizations to be more clear and interpretable.

Age column:

The last change that we made was adding an age column. The data set included the year, month, and day that the patient was born. We were able to design our data analytics script to calculate the age of the patient based on their birth date and today's date. In doing this, we had to change all of the months in the "month" column to the corresponding integer representation for that month so that we could do the calculations. We also had to change the values in the "year" and "date" columns to be actual integers instead of strings holding the integers. Doing all of this allowed us to calculate the age of the patient which was then used for further analysis of what is affecting the total charges a patient is billed.

These changes were essential in helping us to clean our data. With this cleaned data set, we are able to move forward and confidently make data-based decisions and predictions surrounding patients and the charges they will be responsible for.

Model Development:

Objective:

The goal is to predict hospitalization charges using a range of predictors, including patient health metrics (BMI, HBA1C, age), hospital characteristics (hospital tier, city tier), and medical history (heart issues, transplants, cancer history, number of major surgeries, smoking status).

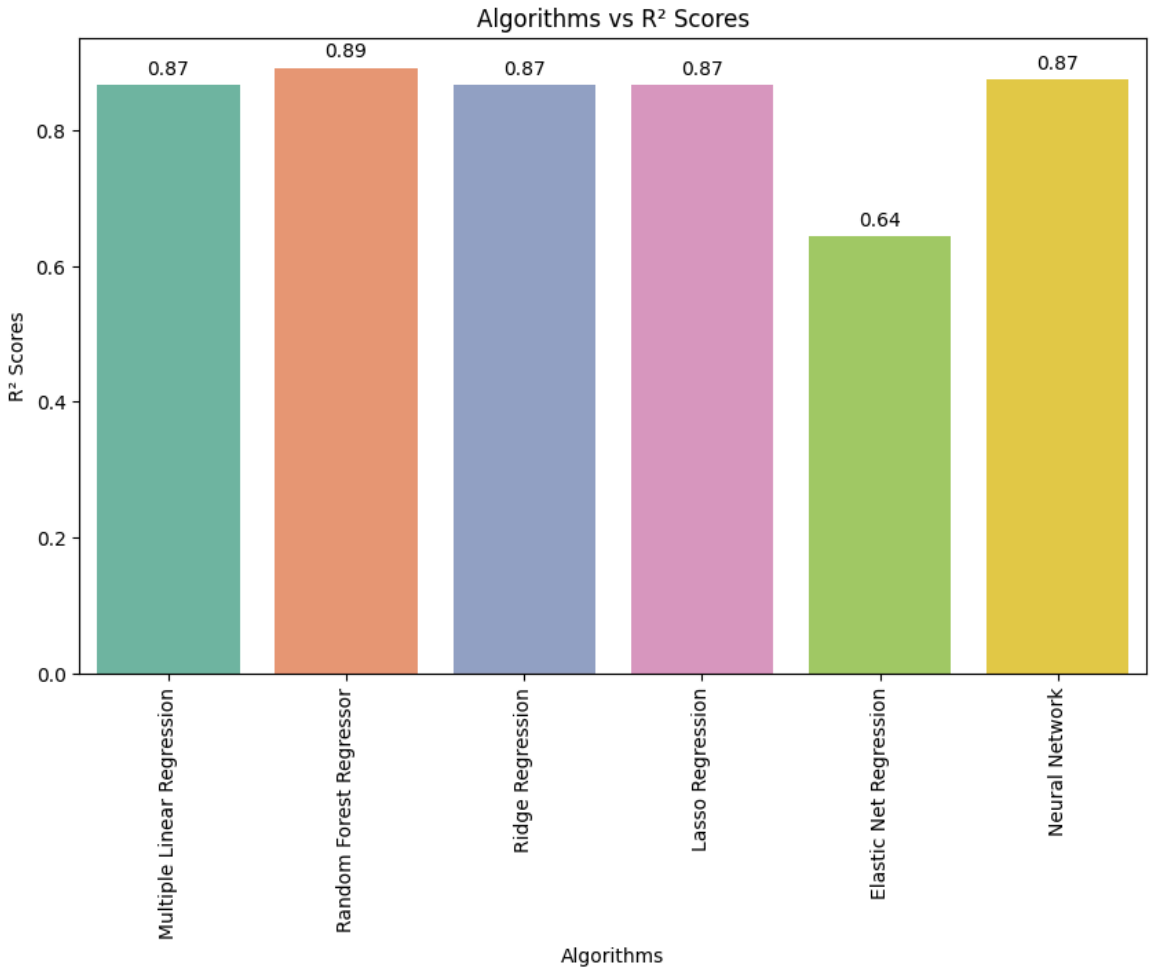
Models Evaluated:

- **Multiple Linear Regression**: Baseline model for linear relationships.
- **Random Forest Regressor**: Non-linear ensemble model.
- **Regularization:** prevents overfitting
    - **Ridge Regression**: Regularized linear regression (L2 penalty).
    - **Lasso Regression**: Regularized linear regression (L1 penalty) for feature selection.
    - **Elastic Net**: Combines L1 and L2 regularization.
- **Neural Network**: Non-linear model capable of learning complex patterns.

Results:

scores and values may differ slightly throughout write-up because the predictions were ran multiple times

Model	R2 Score	RMSE Value	MAE Value
Multiple Linear Regression	0.8760	4442.05	2784.88
Random Forest Regressor	0.9021	3921.35	2465.06
Ridge	0.8762	4443.31	2788.21
Lasso	0.8762	4441.77	2783.06
Elastic Net	0.6598	7246.77	5552.92
Neural Network	0.8812	4316.45	2696.77



Key Findings:

Linear Regression:

- **R^2**: 0.866 — This model explains approximately 86.6% of the variance in hospitalization charges, which is reasonable but not the best.
- **RMSE**: 4442 — The Root Mean Squared Error suggests moderate prediction error.
- **MAE**: 2784 — The Mean Absolute Error indicates the average deviation from the actual charges is around 2784 units.
- Interpretation: Linear Regression provides a good but not perfect fit, as seen in the residual plot. It struggles with larger charges, likely due to its inability to capture nonlinear relationships.

Ridge Regression:

- **R^2**: 0.866 — Ridge Regression yields a similar R^2 to Linear Regression.
- **RMSE** and **MAE** are also close to Linear Regression, showing slight regularization benefits without substantial improvement in fit.
- Interpretation: Ridge helps prevent overfitting by slightly shrinking coefficients but does not significantly outperform Linear Regression.

Random Forest Regressor:

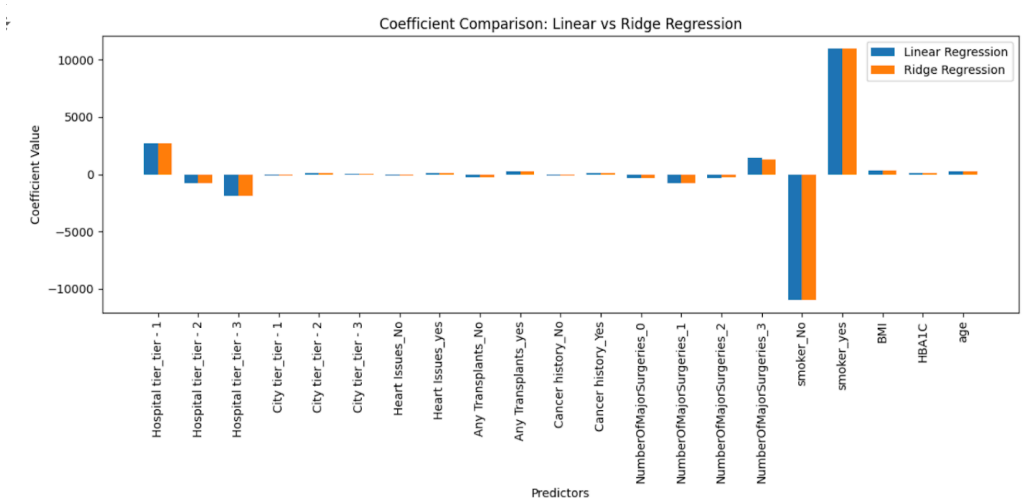
- **R^2**: 0.896 — The Random Forest model explains around 89.6% of the variance, outperforming the linear models, indicating it captures additional patterns in the data.
- **RMSE**: 3921 and **MAE**: 2465 — Both error metrics are lower than those of the linear models, confirming better predictive performance.
- Interpretation: This model's superior performance suggests it benefits from its ability to capture nonlinear relationships and interactions, especially relevant given the dominant impact of smoker status and health metrics.

Neural Networks:

- **Neural Network (R^2: 0.881):** This model performs better than Linear and Ridge, capturing some complex patterns but slightly underperforms Random Forest.
- Interpretation: Neural Networks have a slight edge over linear models by capturing nonlinearities but may lack the interpretability and robustness of Random Forest for this dataset.

Visualizations:

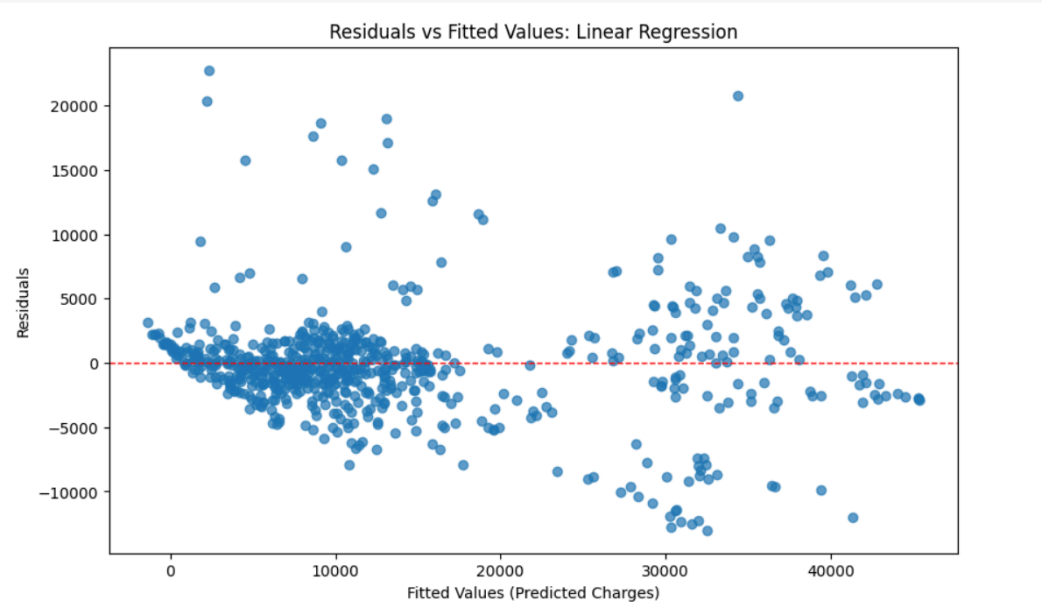
1. **Coefficient Comparison Plot: Linear Regression vs. Ridge Regression**



Insights:

- **Smoker Status** (smoker_yes and smoker_No) has a substantial influence, with smoker_yes having a strong positive effect and smoker_No a strong negative one. This suggests that being a smoker is associated with higher hospitalization charges.
- **BMI, HBA1C, and Age**: These health metrics have smaller coefficients compared to smoker status, indicating a relatively lower impact on charges.
- **Hospital Tier and City Tier**: The coefficients for hospital and city tiers are also small, suggesting that while hospital and city characteristics influence charges, their impact is relatively minor compared to smoking status.
- **Effect of Ridge Regularization**: Ridge regression tends to shrink coefficients slightly toward zero to prevent overfitting. This is noticeable where Ridge coefficients (orange bars) are slightly smaller than the Linear Regression coefficients (blue bars) for some predictors, showing regularization effects.

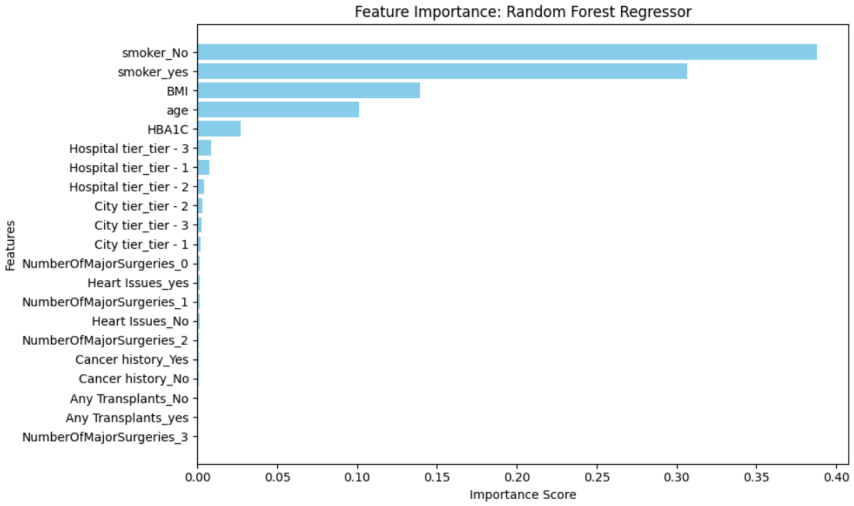
2. **Residuals vs. Fitted Values Plot: Linear Regression**



Insights:

- **Pattern of Residuals**: The spread of residuals increases with predicted charges, which suggests potential heteroscedasticity (i.e., variance of errors changes with the level of the predicted charges).
- **Outliers**: Several points are far from the zero line, especially on the positive side, indicating instances where the model significantly underestimates actual charges.
- **Model Fit**: The residuals are not randomly scattered around zero, suggesting that the Linear Regression model may not fully capture the complexity of the data, likely because of some nonlinear relationships or unaccounted factors influencing charges.

3. **Feature Importance Plot: Random Forest Regressor**



Insights:

- **Dominant Features**:
  - **Smoker Status** (smoker_No and smoker_yes): These two features are by far the most important, indicating that smoking status is the most significant factor in predicting charges. This is consistent with the coefficient plot for linear models.
  - **BMI and Age**: These features also hold significant importance, highlighting that patient health metrics are key predictors of hospitalization charges.
  - **Other Features**: Hospital and city tiers, as well as medical history variables (e.g., number of major surgeries, heart issues), have lower importance scores, indicating they are less influential in this dataset.
- **Nonlinearity and Interactions**: The Random Forest model's ability to capture non-linear relationships and interactions among features may explain its higher performance metrics.

Key Interpretations:

Key Predictors: Smoker status, BMI, and age are the most influential features in predicting hospitalization charges. This is consistent across different models, with both linear and non-linear models highlighting these features.

Best Overall Model: The Random Forest Regressor performs the best with the highest R^2 and lowest errors, suggesting it's the most suitable model for predicting hospitalization charges due to its capability to model non-linearities and interactions.

Trade-offs with Linear Models: Linear and Ridge Regression provide interpretable results with reasonable accuracy, useful for understanding the direction and strength of each predictor's effect. However, they fail to capture complex patterns, particularly in high-charge cases.

Residual Analysis:

Linear models showed a well-distributed residual pattern but lacked the flexibility to capture non-linearities.

Non-linear models such as Random Forest demonstrated superior fit, addressing these patterns effectively.

Practical Implications and Recommendations:

Smoker Status as a Primary Driver: The models consistently highlight smoker status (smoker_yes and smoker_No) as the most significant predictor of hospitalization charges. This aligns with the increased healthcare needs of smokers due to conditions like respiratory issues, cardiovascular diseases, and cancer, which are often costly to treat.

Health Metrics (BMI, Age, HBA1C): BMI and age also significantly influence costs. Higher BMI is associated with obesity-related health complications, and age is often correlated with chronic conditions and increased healthcare utilization. HBA1C (indicative of diabetes) has a smaller impact but remains a meaningful predictor.

Low Impact of Hospital and City Tiers: The analysis reveals that hospital and city tiers have relatively low predictive importance for hospitalization charges. This suggests that geographic or institutional factors may not strongly affect the costs compared to patient health and lifestyle factors.

Random Forest Model for Cost Prediction: Given the Random Forest model's superior accuracy, it should be used for operational predictions of hospitalization charges. Its ability to

capture nonlinear relationships and interactions among features makes it suitable for accurately predicting costs, especially for complex cases.

Pattern of Residuals Suggests Uncaptured Complexity: The residuals analysis for Linear Regression suggests heteroscedasticity, indicating that the variability in hospitalization costs increases with the predicted amount. This may imply that other unobserved factors influence costs, especially at the high end.

Addressing the Problem

The original problem highlights the challenges faced by healthcare organizations in forecasting hospitalization costs and the associated impacts on financial planning, resource utilization, and patient stress. The insights gained from the analysis directly address these issues:

1. Improved Cost Forecasting:

The analysis identifies key predictors of hospitalization charges, such as smoking status, BMI, and age. By focusing on these variables, healthcare organizations can develop more accurate cost models, reducing budget overruns and resource misallocation.

2. Transparent Estimates for Patients:

The findings enable healthcare providers to offer more personalized and precise cost estimates to patients, addressing financial uncertainties and reducing the likelihood of delayed or skipped care due to cost concerns.

3. Resource Allocation:

Insights into the low impact of hospital and city tiers allow organizations to focus on patient-centric factors rather than institutional adjustments. This ensures resources are allocated where they have the most significant effect on cost optimization.

4. Health Initiatives to Lower Costs:

Highlighting the high costs associated with smoking and obesity provides actionable evidence to support targeted health programs. By addressing these factors, organizations can proactively reduce healthcare expenses.

Actionable Recommendations

1. Develop a Patient-Centric Cost Estimation Tool:

Create a digital platform or dashboard using the Random Forest model to provide patients with personalized cost estimates based on their health metrics and medical history. This transparency can help improve trust and satisfaction.

2. Implement Targeted Health Programs:

Focus on smoking cessation and obesity management programs. These initiatives can reduce the prevalence of high-cost cases, directly addressing the significant predictors identified in the analysis.

3. Enhance Data Collection:

Collect additional patient data, such as socio-economic factors, treatment intensity, and satisfaction levels, to further improve model accuracy and capture unexplained variability in high-cost cases.

4. Adopt the Random Forest Model for Operational Use:

Train healthcare administrators and finance teams to use the Random Forest model for regular cost forecasting. Its non-linear capabilities make it ideal for handling complex patient profiles.

5. Adjust Pricing and Financial Planning:

Utilize the model's predictions to set more realistic financial targets and pricing strategies. This will improve financial stability and allow for proactive adjustments in resource planning.

6. Monitor Residual Patterns for Continuous Improvement:

Regularly analyze residual patterns to identify potential gaps in the model. This ensures continuous refinement and adaptation to evolving healthcare trends.

7. Expand Preventive Care Strategies:

Use the model to identify high-risk patient groups and develop preventive care plans, such as regular health screenings for smokers or tailored interventions for older patients with high BMI.

Conclusion

Summary:
This project successfully developed predictive models for hospitalization charges using a comprehensive set of patient and hospital-level predictors. Key findings highlight the dominant role of patient health metrics—especially smoker status, BMI, and age—in determining costs. The Random Forest Regressor emerged as the most suitable model due to its ability to handle non-linear relationships and interactions. Insights gained can guide healthcare organizations in financial planning, resource allocation, and improving patient affordability.

Impact:
The results address a critical issue for healthcare organizations by providing a framework for more accurate cost forecasting, which could lead to better financial management and enhanced patient care. Transparent cost estimates can empower patients to plan their expenses, reducing the stress and barriers to accessing necessary care.

Acknowledgements:

We would like to sincerely thank Prof. Michael Rivera for teaching us everything we needed to know and for going above and beyond to keep us engaged. His dedication and enthusiasm made learning enjoyable and helped us truly understand and retain the material.

We're also grateful to the Lehigh College of Business for all the resources and support throughout this project. A big shout out to the amazing team—Josh Garcia, Wei Fang, and Milonee Parekh—for their hard work and collaboration in making this project a success.