PREDICTING PATIENT READMISSIONS IN HEALTHCARE

BUS 9430 - Business Analytics Project Management

TEAM: MediForecast Consultants

Rupa Poddar, Sahithi Arnika Modadugu,Sujasna Tamang,
Mustafa Ekinci.

Introduction / Research Questions

Predictive modeling for readmission risk assessment: Can we accurately predict patient readmission based on attributes such as medical history, admission details, and treatment plans?

Models

Logistic Regression

Support Vector Machine (SVM)

Feature Selection

Used Random Forest Classifier for feature selection

Selected feature:

Number of lab procedures

Number of medications

Time in hospital

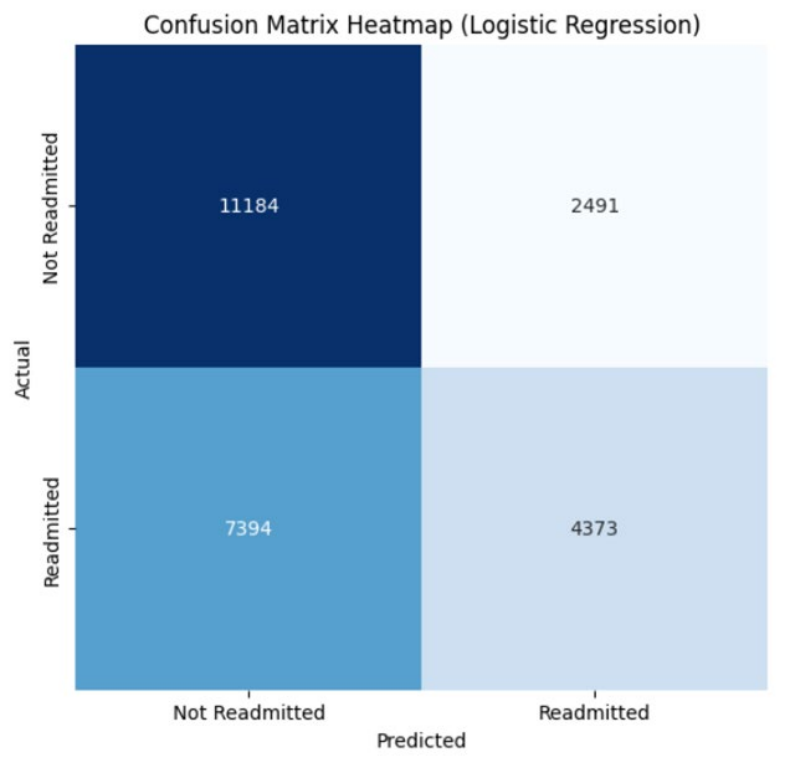
Number of diagnoses

Number of inpatient Visits

Logistic Regression Model

Confusion Matrix Overview:
True Negatives (TN): 11,184 - Correctly predicted non-readmission.
False Positives (FP): 2,491 - Incorrectly predicted as readmission.
False Negatives (FN): 7,394 - Incorrectly predicted as non-readmission.
True Positives (TP): 4,373 - Correctly predicted readmission.

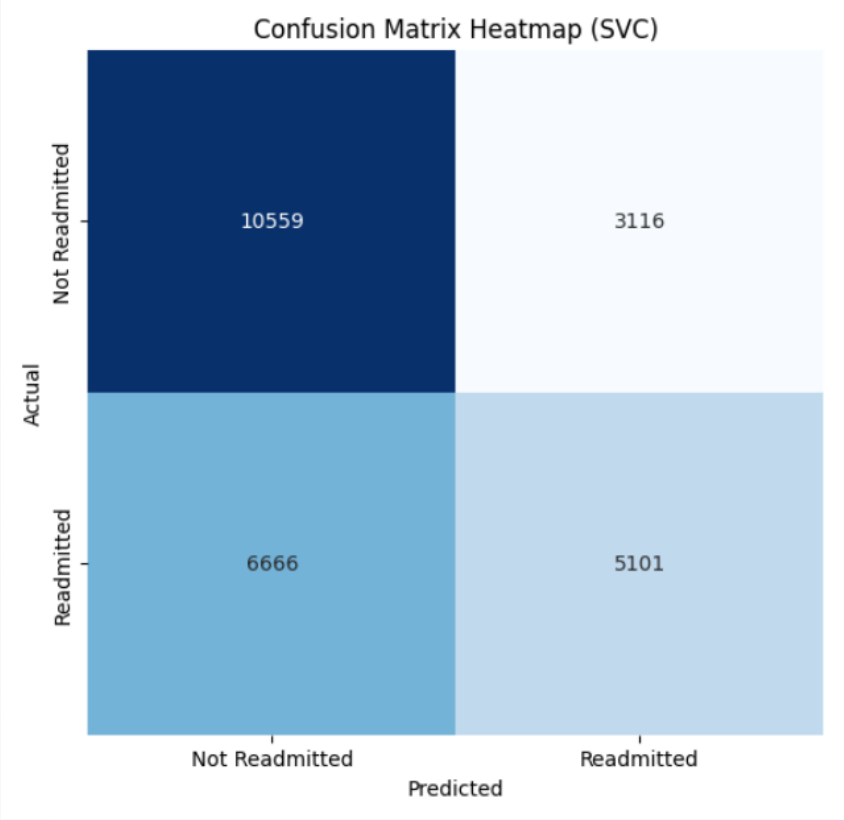
Accuracy:
Result: ~61.1% - Overall, 61.1% of predictions were correct.



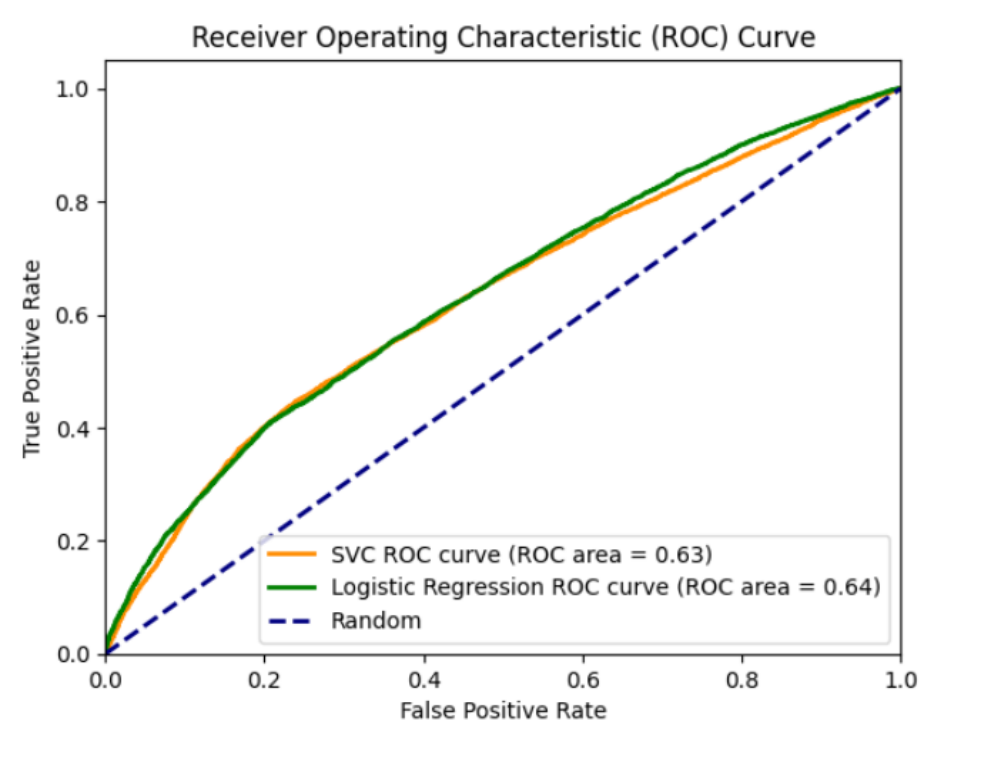
Support Vector Machine Model

Confusion Matrix Overview:
True Negatives (TN): 10, 559 - Correctly predicted non-readmission.
False Positives (FP): 3, 116 - Incorrectly predicted as readmission.
False Negatives (FN): 6, 666 - Incorrectly predicted as non-readmission.
True Positives (TP): 5, 101- Correctly predicted readmission.

Accuracy:
Result: ~62% - Overall, 62% of predictions were correct.



ROC Curve



No Significant difference between SVC and Logistic regression in terms of ROC area.

Cost Benefit Estimation

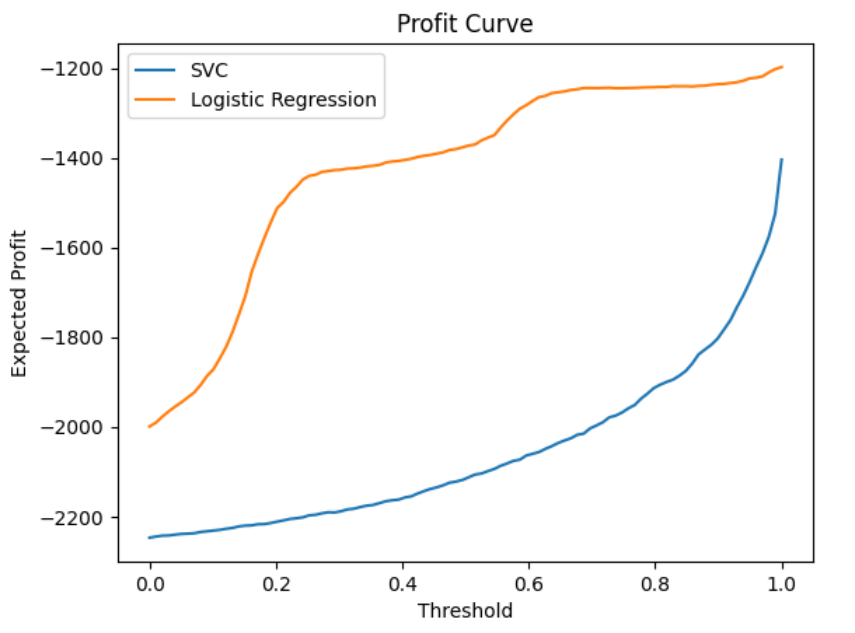
True Positive: True Positive might not incur any direct benefit or cost - $0

False Negative: Estimated additional cost due to potential complications - ($13000)

False Positive: Opportunity cost - ($2,000)

True Negative: Avoidance of unnecessary interventions - $0

Profit Curve



Logistic Regression:
Significant losses but demonstrates a gradual increase in profit as the decision threshold increases, peaking around the mid-range threshold.
The threshold approaches 1, there is a sharp decline in profit, indicating diminishing returns at higher thresholds.

SVC:
Negative profit across all thresholds.
gradual improvement noted as the threshold increases; yet, it remains unprofitable throughout the entire range, never crossing into positive profit territory.

Interpretation

- Comparing accuracy or ROC curve did not help in determining the best model.
- Logistic Regression is suitable to predict if the patients will be readmitted or not based on the profit curve.

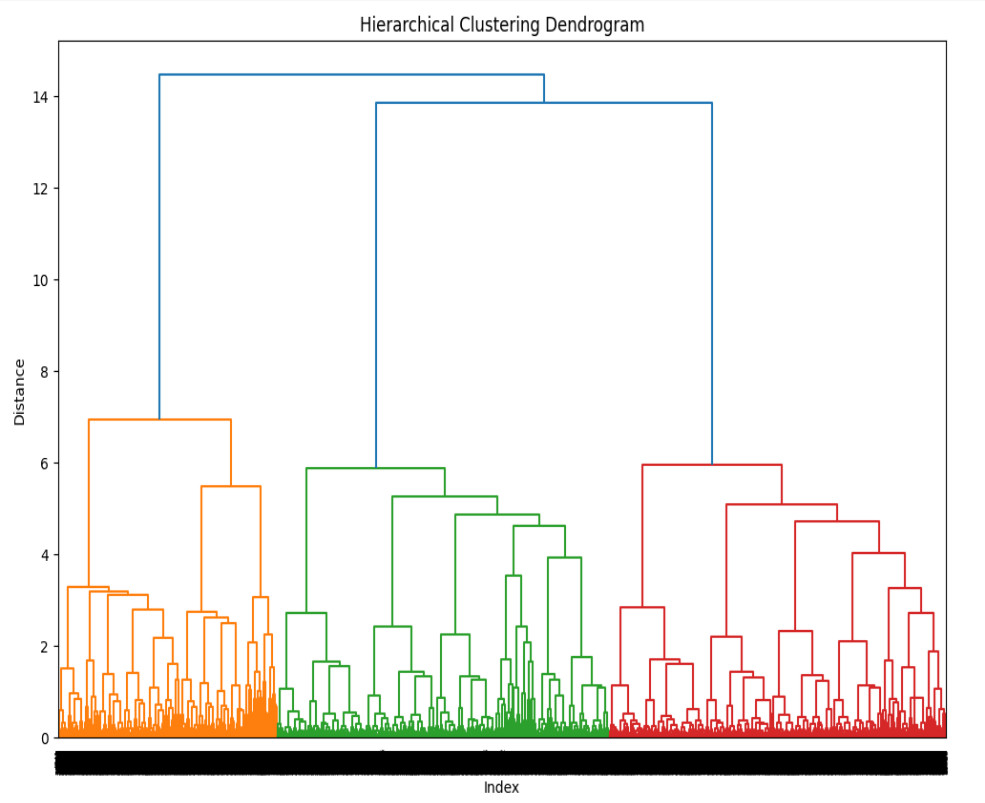
Research Question 2

Segmentation of patients based on risk factors: Can we segment patients into different risk groups based on their medical profiles and demographic characteristics?

Models

- Hierarchical clustering
- K-Means Clustering

Hierarchical Clustering



- Utilized hierarchical clustering to uncover underlying patterns in patient data using both numerical and categorical data like race, age and gender.
- The dendrogram illustrates the hierarchical relationships between data points based on their similarity.
- Each leaf node in the dendrogram represents an individual data point, while internal nodes represent clusters of data points.
- Three major clusters were formed and we have grouped patients by cluster and calculated the average readmission rate for each cluster.

Cluster Analysis: Readmission Rates by Patient Cluster

Cluster	Readmission rate
1	41.7%.
2	39.4%
3	38.4%

The calculated readmission rates for each cluster show no significant differences, indicating similar trends across clusters.

Silhouette Score:

- The silhouette score measures the compactness and separation of clusters in the clustering analysis.
- Silhouette score ranges from -1 to 1: Closer to 1 means well-separated clusters, 0 implies overlapping clusters, and negative values indicate potential mis assignments.
- In this analysis, the silhouette score is approximately **0.0566**, indicating a modest degree of separation between clusters.

K-Means Clustering (with K=3)

Cluster	Readmission rate	
1	52.6%.	High- Risk Patient
2	40.8%	Medium -Risk Patient
3	32.7%	Low-Risk Patient

The calculated readmission rates for each cluster show significant differences across clusters. We can cluster these 3 segment patients as High-Risk, Medium-Risk and Low-Risk

Silhouette Score:

In this analysis, the silhouette score is approximately **0.1592**, indicating a better degree of separation between clusters than Hierarchical clustering

Interpretation

K-Means Advantage: K-means clustering is more effective for our data, providing clear segmentation of patient readmission risk and better cluster separation.

Hierarchical Clustering Limitation: The hierarchical approach did not yield distinct clusters, indicating potential issues with cluster separation or data suitability for this method

Conclusion

“Hospital Management can use models like logistic regression and K-Means clustering for predicting whether the patient will be readmitted or not and segmenting patients into clusters. This allows them to estimate their risk of readmission, providing valuable insights for targeted interventions and personalized healthcare strategies”.

THANK YOU!