WINONA
STATE UNIVERSITY

Using Medical Data to Predict Future Patient Expenditures

Will Diedrick, Jimmy Hickey, Akif Khan,
Kapil Khanal, Sean Wittenberg

Minne Analytics

Introduction

- An employer is seeking to save money by helping their employees with type 2 diabetes to lead healthier lifestyles.
- The employer understands it will be more efficient to cut costs by focusing on those who will be high cost in the future.
- The employer has provided a large data set of persons with type 2 diabetes.
- Our team has been presented with the challenge of analyzing the data.

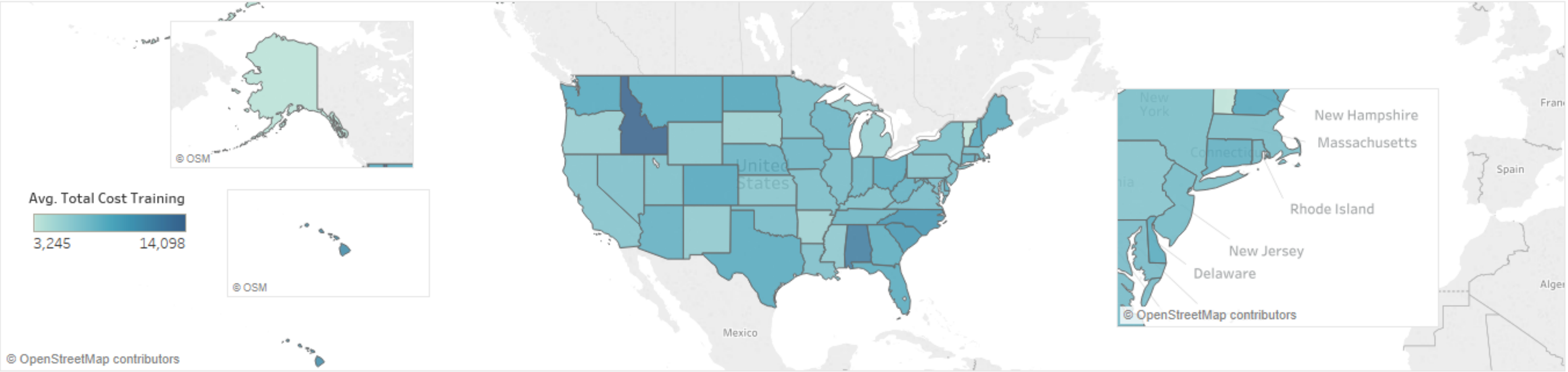
Objective

- Organize a large data set to allow for a more manageable investigation.
- Examine key differences between low and high cost patients.
- Make meaningful insights.
- Tell a story with the data.

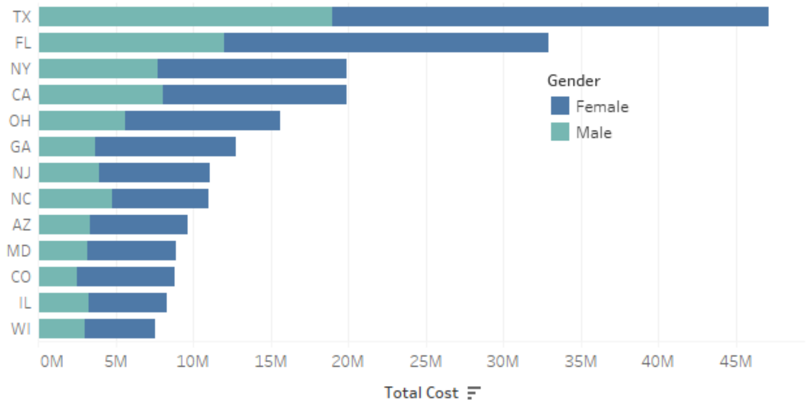
Managing the Data

- Incorporate lookup tables to create meaningful variables.
- Adjust the training data set to the scale of the target data set.
- Merge target and training data set.
- Collapse each patient to a single row.
- Rank the patients by their total cost.

Average Patient Cost per State

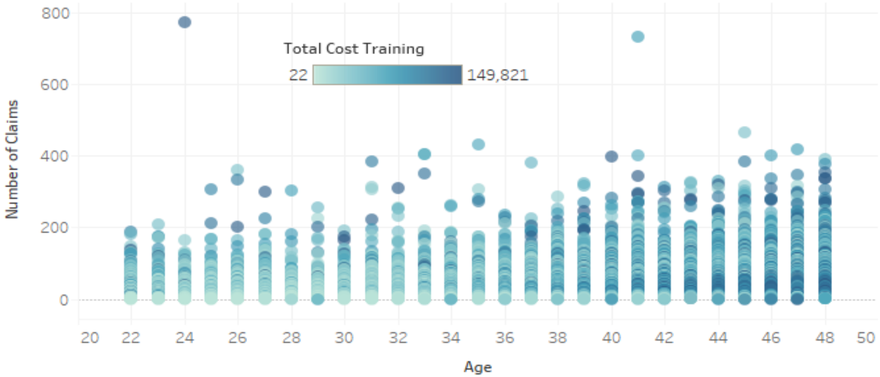


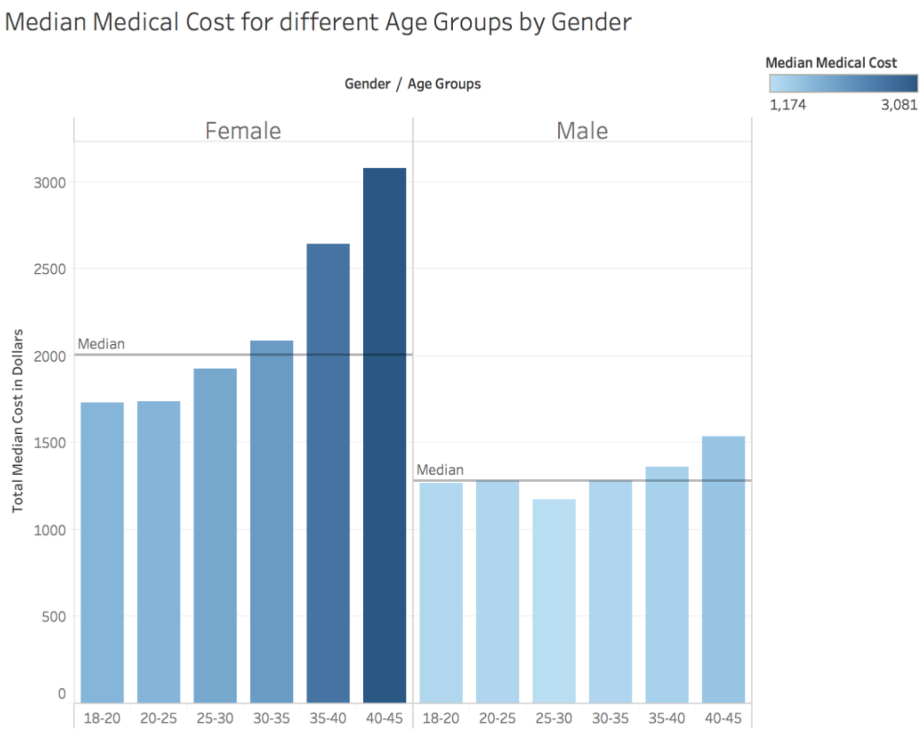
Total Cost of States Split by Gender



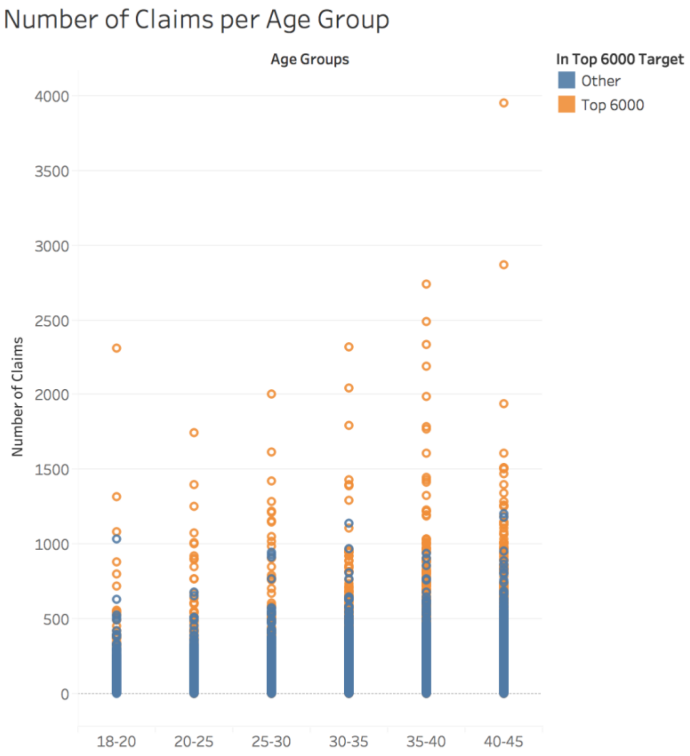
Age, Number of Claims, and Total Cost of Patients

Total cost is represented by the color of each dot

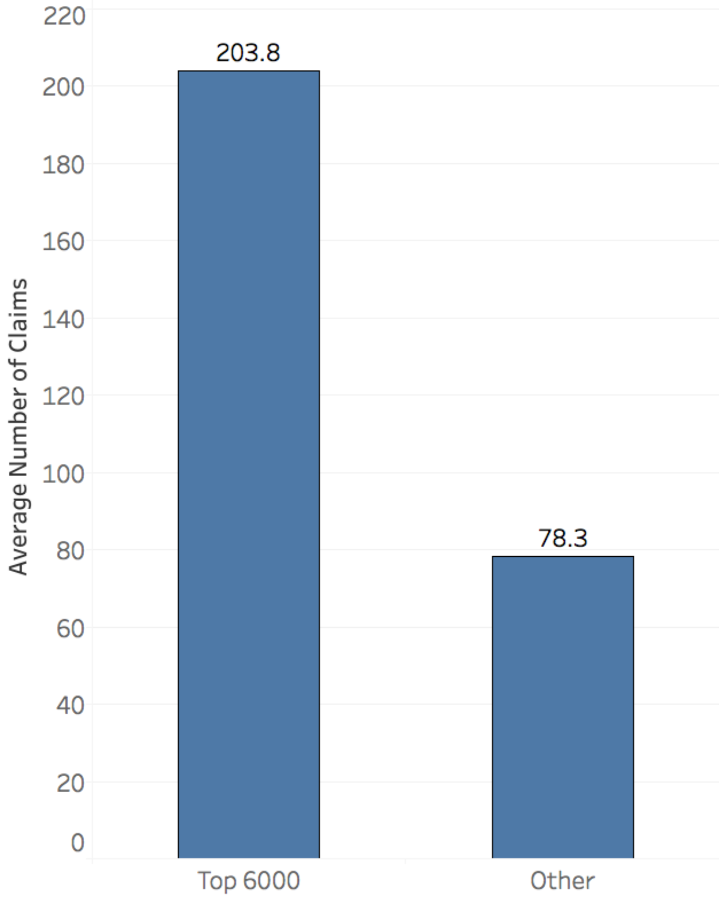




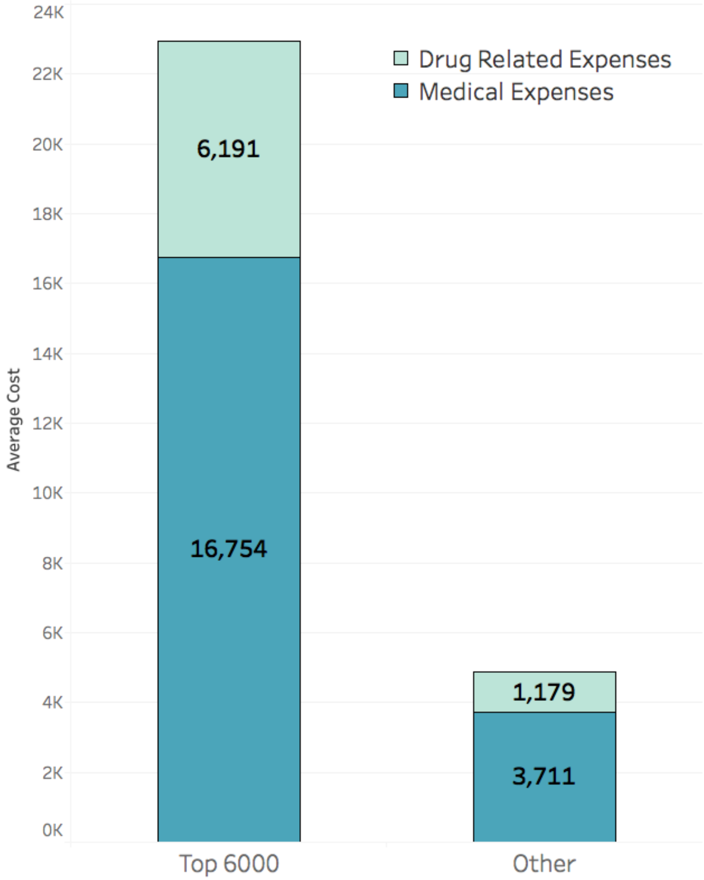
Median of Total Cost Medical for each Age Groups broken down by Gender. Color shows median of Total Cost Medical. The view is filtered on Gender, which keeps Female and Male.

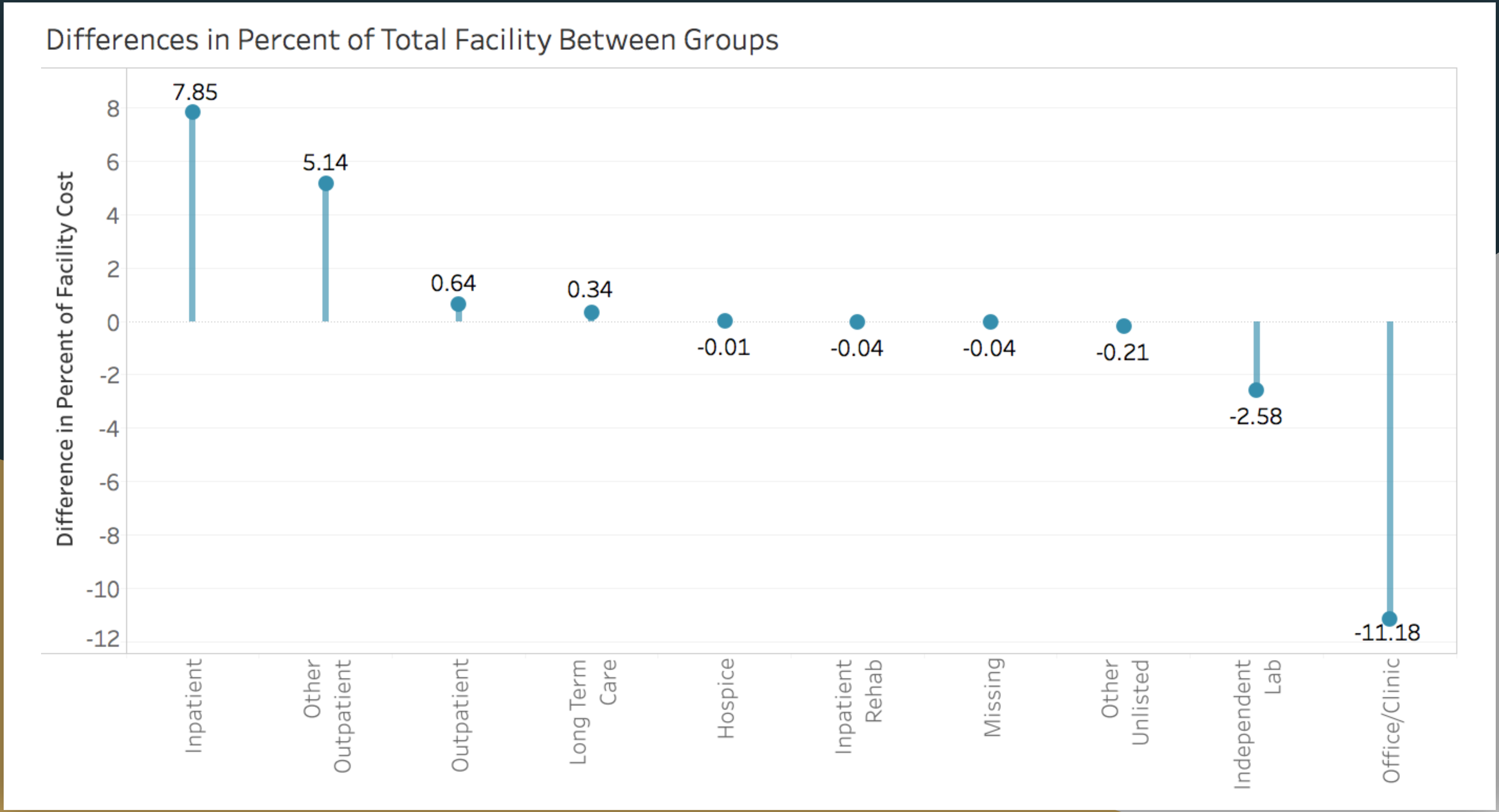


Number of Claims Across Groups per Year



Total Cost Broken Down Across Groups




Differences in Percent of Total Facility Between Groups
Difference in Percent of Facility Cost
8
6
4
2
0
-2
-4
-6
-8
-10
-12
7.85
5.14
0.64
0.34
-0.01
-0.04
-0.04
-0.21
-2.58
-11.18
Inpatient
Other Outpatient
Outpatient
Long Term Care
Hospice
Inpatient Rehab
Missing
Other Unlisted
Independent Lab
Office/Clinic

Modeling and Analysis

Predicting and classifying high cost patients

Response: Top 6000 Patients in Target Set(Next Year)

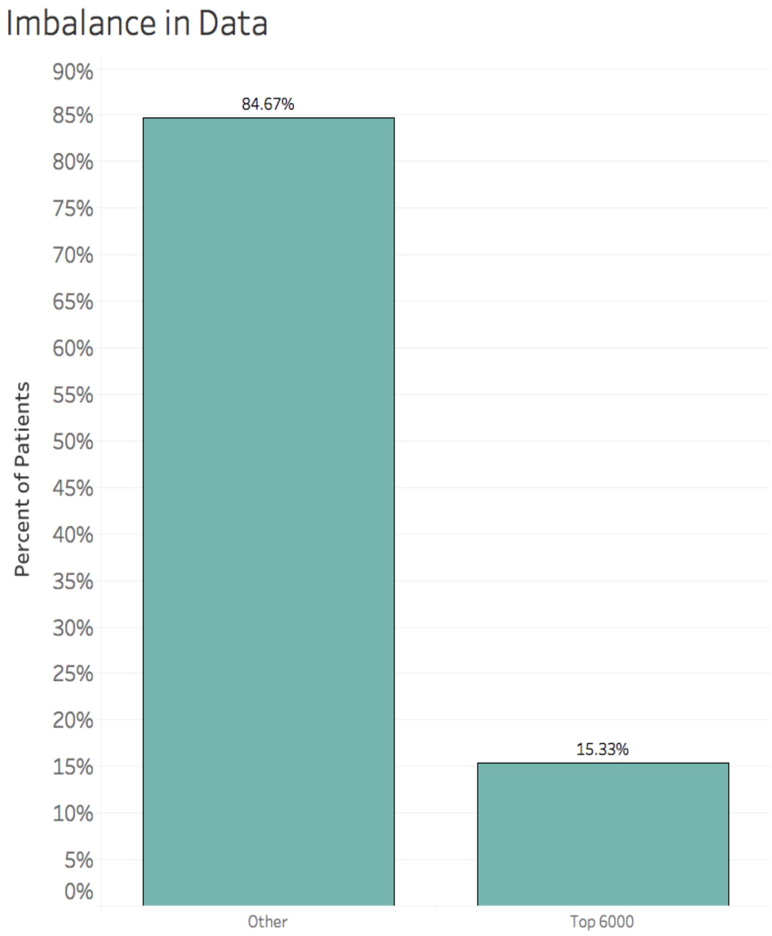
Predictors: Medical Data from Training Set(Current Year)

Data Distribution:

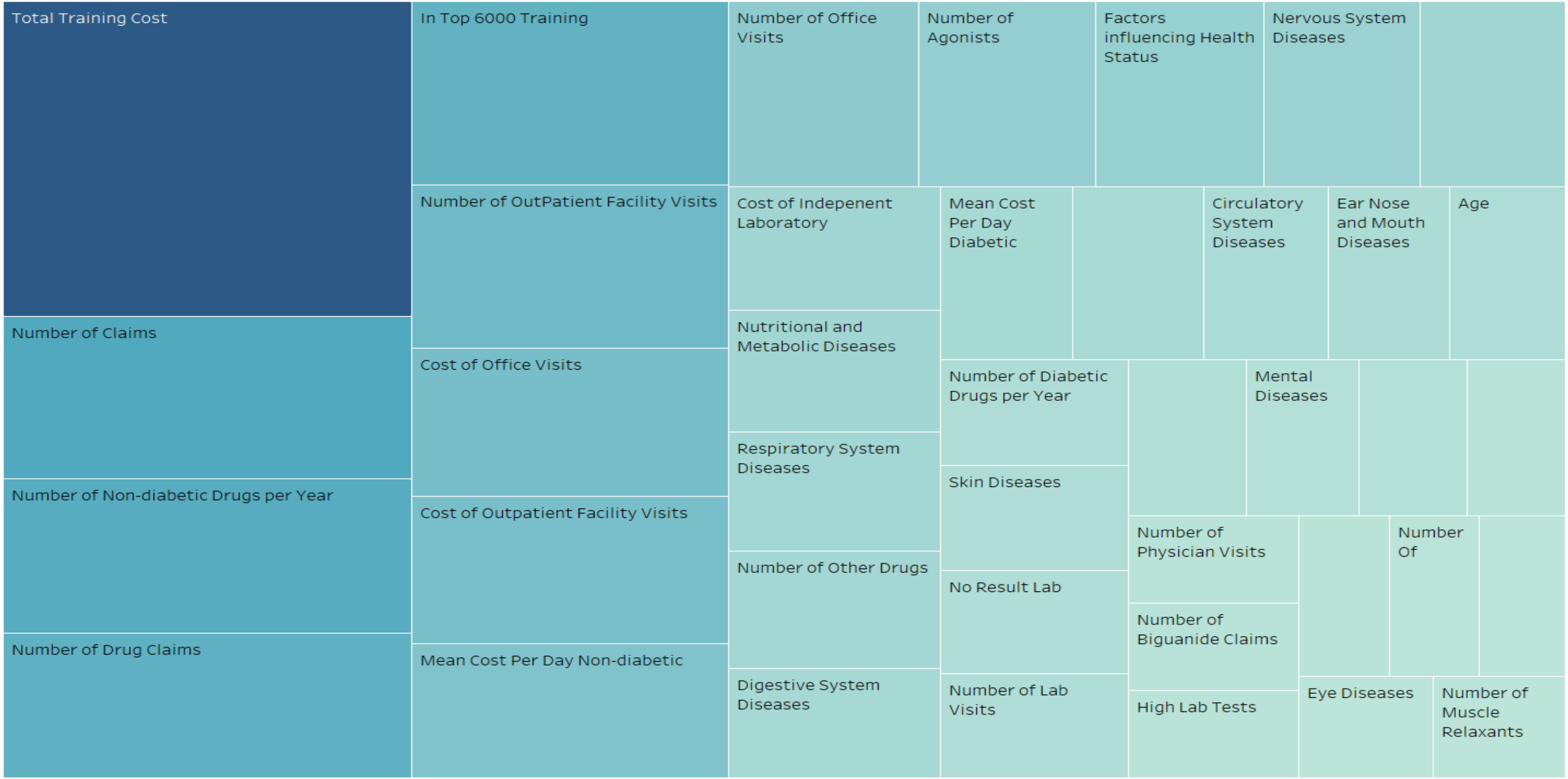
The data is highly unbalanced making the algorithms biased towards the majority classes in the predictors.Thus, we used sampling techniques to balance the classes in the data

Feature Engineering:

We used a Random Forest Model to assess the most significant predictors.



Variables that are Most Explanatory of Top 6000

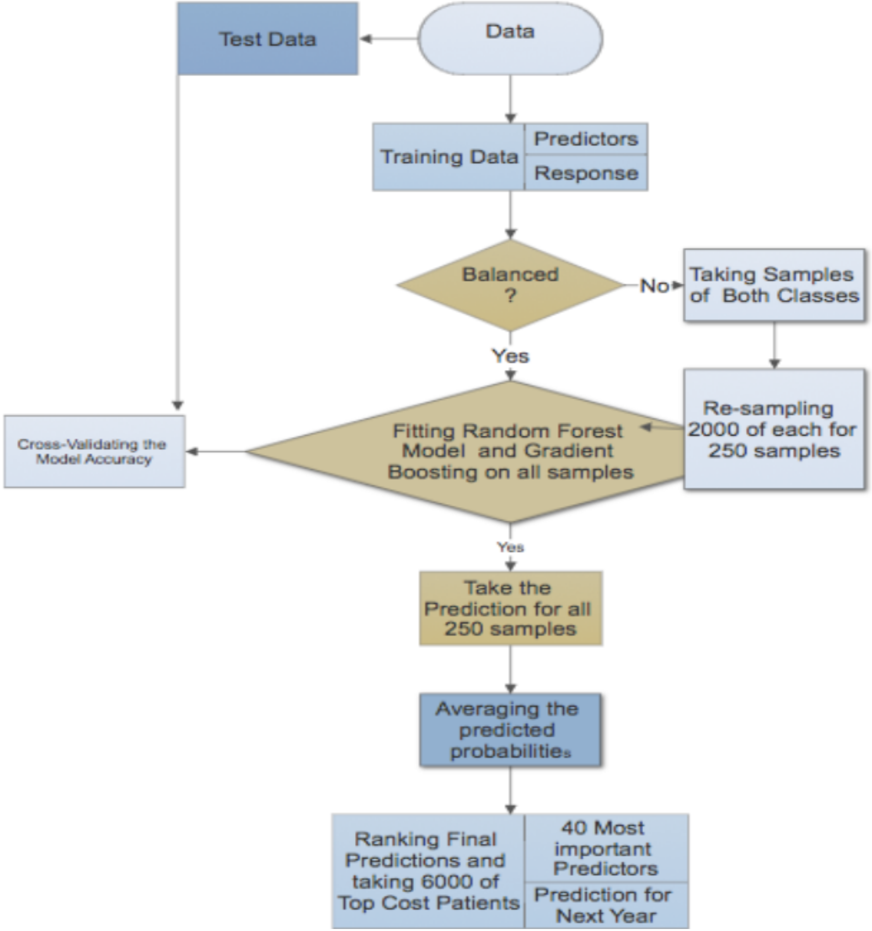


FlowChart of algorithm used for the modeling.

Repeated sampling to balance the data and have decrease its effect.

The Random Forest helps decrease the variance in the data while gradient boosting classifiers to decrease the bias.

Model properly predicts around 55% of 6000 patients.



Conclusion

- The predictions from our model will help the employer understand the characteristics of diabetic employees that are likely to be high cost in the future.
- This insight will help allow the employer to identify these employees and focus on improving the health of these patients.
- Successfully identifying these employees will both reduce costs for the employer but also benefits the potentially high cost employees.