INSURANCE CONSULTING

Shehzad Ali, Rianna Patel, Qingye Ding, Yashpreet Kaur, and Trevor Moos

TABLE OF CONTENTS

PROBLEM STATEMENT

What are we looking to solve?

01

EXPLORATORY DATA ANALYSIS

Takeaways from our initial investigation

02

MODEL RESULTS

Which model produces the best results?

03

CONCLUSION

How does our analysis help the company?

04

01

UNDERSTANDING THE PROBLEM

What attributes are common in individuals who have high medical cost?

UNDERSTANDING THE PROBLEM

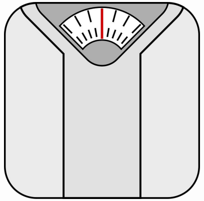
PROBLEM STATEMENT

Can we identify individuals who will have high medical costs based off human variables?

REAL WORLD APPLICATION

By identifying high cost individuals, we can accurately price insurance plans.

VARIABLES

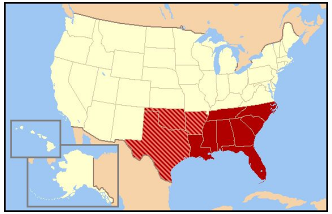


BODY MASS INDEX

CDC categorization: underweight, healthy, overweight, and obese

AGE

Categorized: 18-29, 30-39, 40-49, 50-59, and 60-65

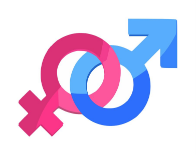


REGION

Categorized: Northeast, Northwest, Southeast, and Southwest

NUMBER OF CHILDREN

Categorized: 0, 1, 2, 3, 4, and 5



SEX

Categorized: Male and Female

SMOKER

Categorized: smoker and non-smoker

MEDICAL CHARGES

Categorical variable: normal and high-cost

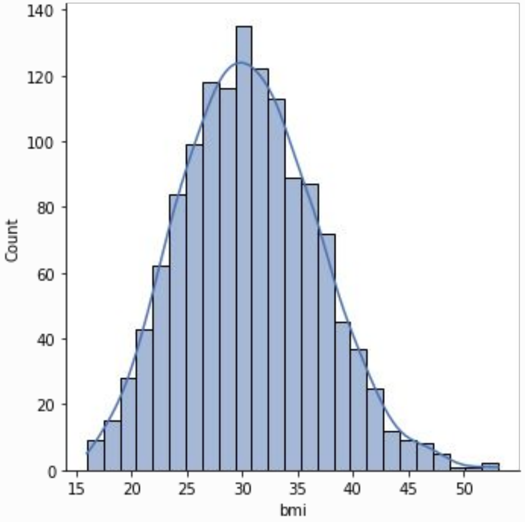
02

EXPLORATORY DATA ANALYSIS

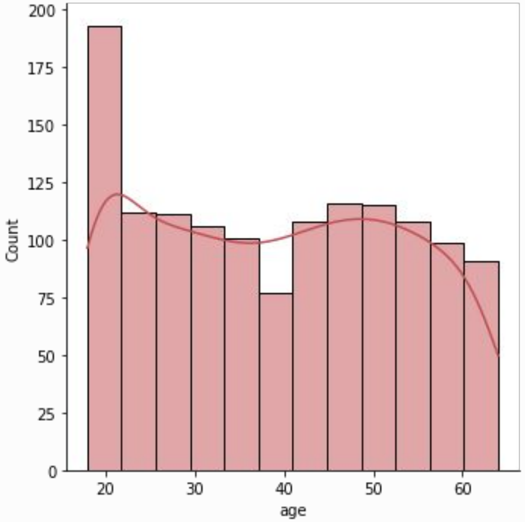
Takeaways from the initial investigation

DATA DISTRIBUTION

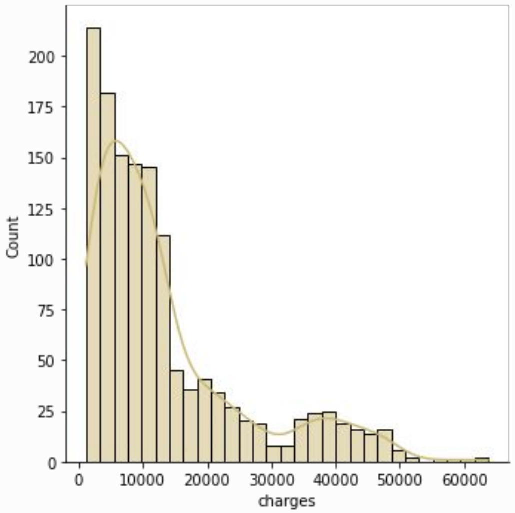
BODY MASS INDEX



AGE



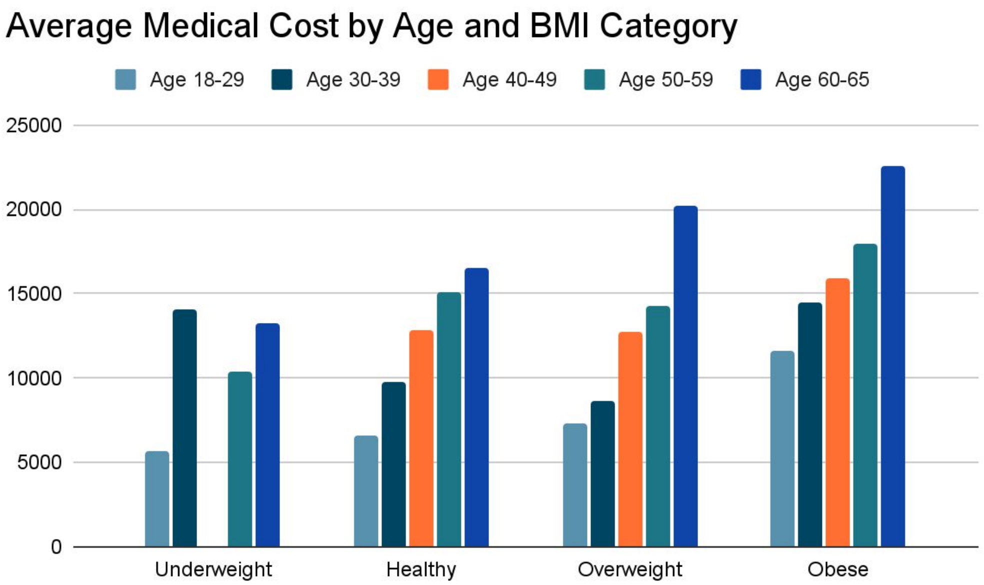
CHARGES



IMPACT OF AGE AND BMI

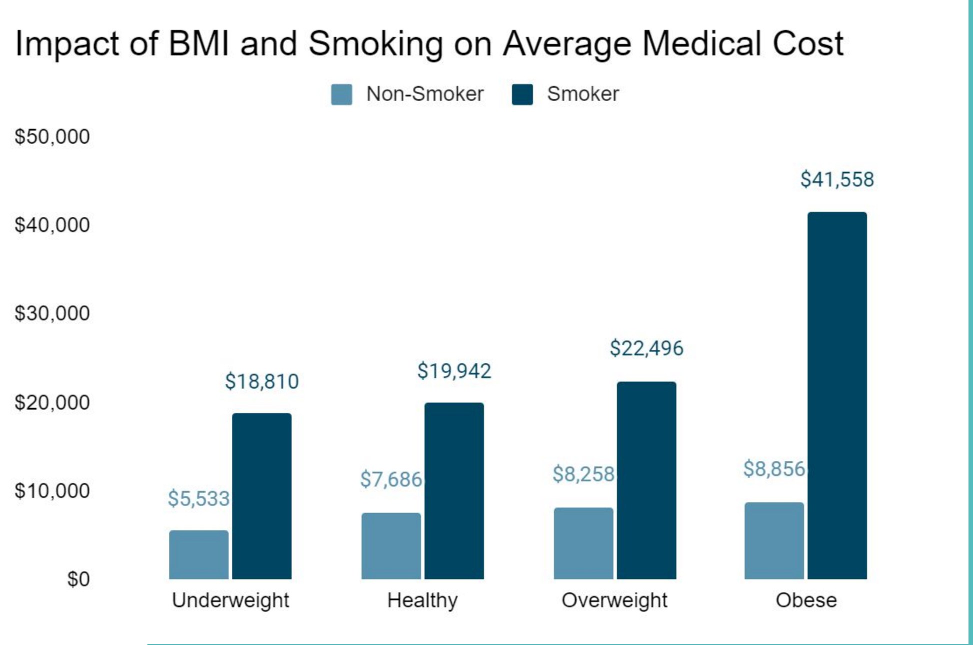
Age Category	Average Medical Cost
18-29	$9,200.62
30-39	$11,738.78
40-49	$14,399.20
50-59	$16,495.23
60-65	$21,248.02

BMI Category	Average Medical Cost
Underweight	$8,852.20
Healthy	$10,987.51
Overweight	$10,409.34
Obese	$15,572.04



IMPACT OF SMOKING

- ❑ Smokers make up 20.5% of the dataset and have 4 times the average medical cost of a non-smoker
- ❑ Being a smoker is the most impactful variable on medical cost



03

PREDICTIVE MODELING

Predicting expected medical charges and identifying high risk individuals

LINEAR REGRESSION

ALL VARIABLES

- ❑ Adjusted R^2: 0.753
- ❑ Significant Variables:
  - ❑ Smoker[Yes]
  - ❑ Age
  - ❑ Body Mass Index
  - ❑ Children
- ❑ AIC: 18960.68

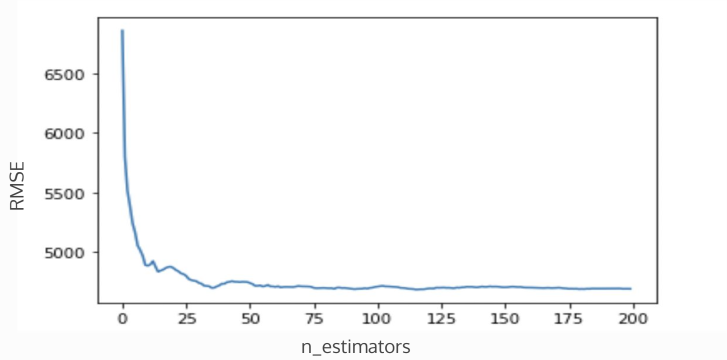
SIGNIFICANT VARIABLES ONLY -- BEST PERFORMANCE

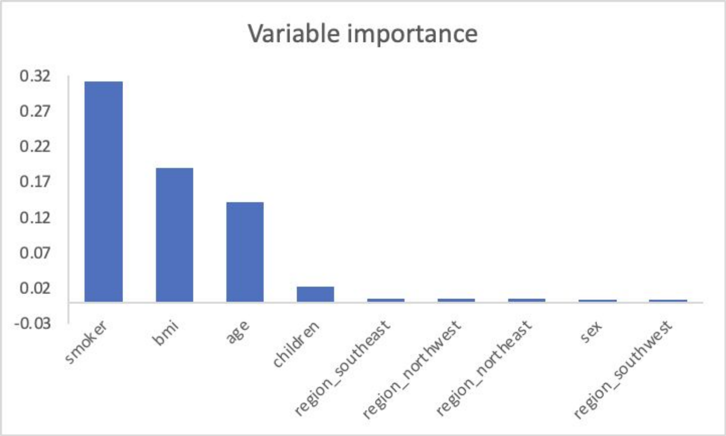
- ❑ Adjusted R^2: 0.752
- ❑ AIC: 18959.63
- ❑ RMSE: 6,253

FEWER SIGNIFICANT VARIABLES

- ❑ Adjusted R^2: 0.749
- ❑ AIC:18968.59

RANDOM FOREST REGRESSION





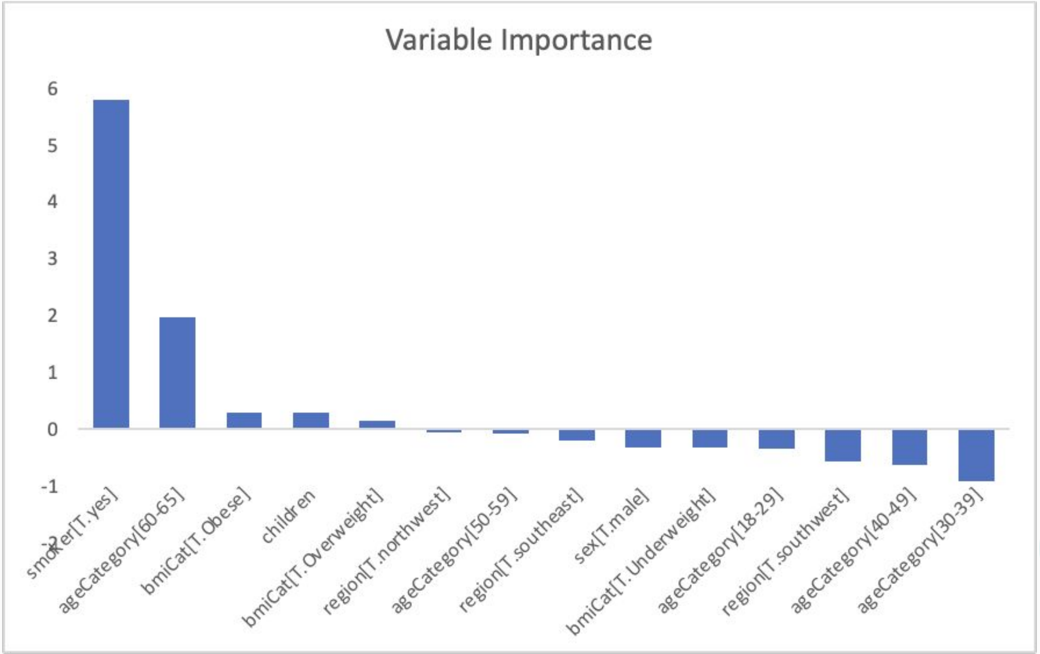
MODEL ACCURACY

- ❏ Min RMSE: 4,689
- ❏ # Trees: 115
- ❏ Important Variables:
  - ❏ Smoker
  - ❏ BMI
  - ❏ Age

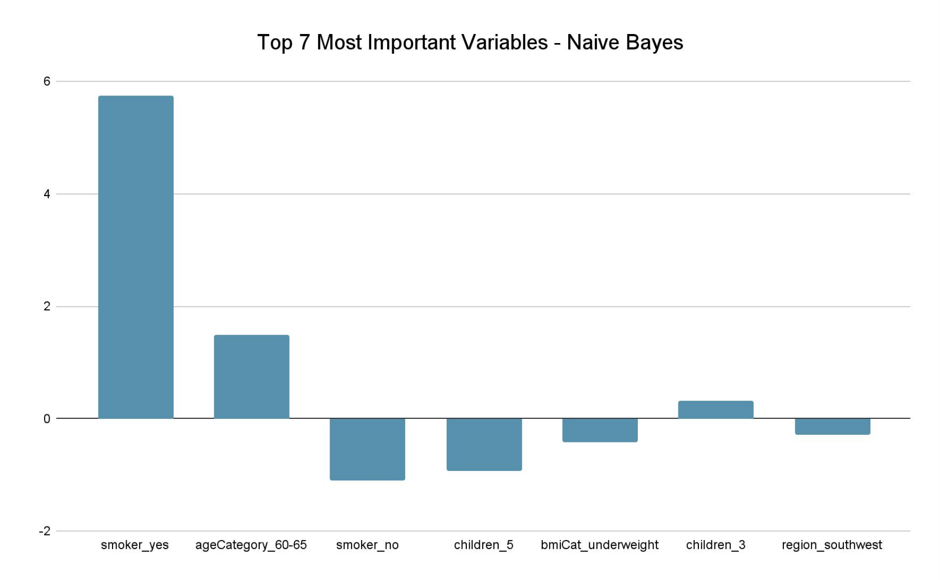
LOGISTIC REGRESSION

MODEL TAKEAWAYS

- ❑ Priors
  - ❑ High cost: 408 (30%)
  - ❑ Normal cost: 929 (70%)
- ❑ Test Accuracy: 92.5%
- ❑ Important Variables:
  - ❑ Smoker[Yes]
  - ❑ Age[60-65]
  - ❑ BMI[Obese]



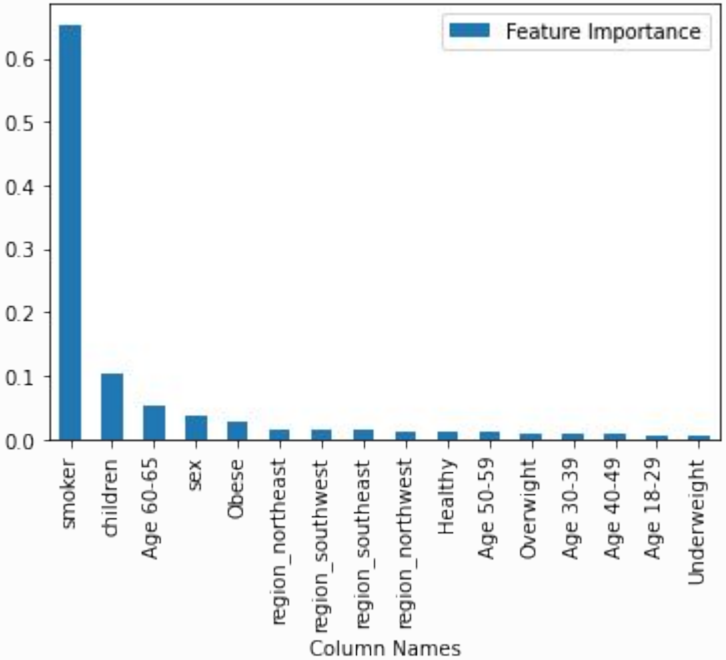
NAIVE BAYES CLASSIFIER



MODEL TAKEAWAYS

- ❑ Test Accuracy: 90.7%
- ❑ Important Variables:
  - ❑ Smoker[Yes]
  - ❑ Age Category[60-65]
  - ❑ Children[5]

RANDOM FOREST CLASSIFIER



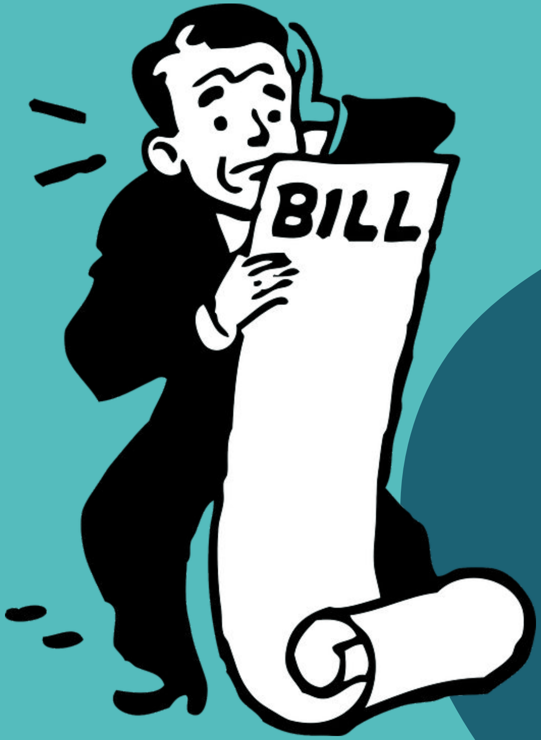
MODEL ACCURACY

- ❑ Test accuracy: 89.9%
- ❑ Important Variables:
    - ❑ Smoker, overwhelmingly so
    - ❑ Children
    - ❑ Age 60-65

04

CONCLUSION AND RECOMMENDATION

How does our analysis impact the business?



CONCLUSION

BEST REGRESSION

RMSE

4,689

RANDOM FOREST REGRESSION

BEST CLASSIFIER

92.5% ACCURACY

LOGISTIC REGRESSION

APPLICATION

- ❑ Smoking is by far the most important factor in predicting medical cost
- ❑ Being able to accurately predict expected medical cost for individuals, will allow us to accurately price our insurance packages.
- ❑ The typical high cost person is 60-65, Obese, and a smoker

THANKS!

ANY QUESTIONS