BMW USED CAR SALES BASELINE PRICE PREDICTION

NICK BULTMAN

OUTLINE

- Problem
- Data
- Analysis
- Results
- Recommendations/Next Steps

PRICING ANALYST PROBLEM – 2-FOLD

- Pricing Cars = Difficult!
  - Accuracy – many different aspects to keep in mind
    - Too high/low? Not maximizing profitability
  - Efficiency – pricing process takes time
    - Longer = more expenses = less profit
- Can data science be used to not only generate a more accurate *baseline* price with historical data available but also speed up the pricing process to better ensure company profitability?

HISTORICAL DATA

- Obtained from GitHub – posted by colleague
- 9 columns, 10,000+ observations of BMW Used Car Sales
  - BMW Model
  - Year
  - Price
  - Transmission Type
  - Mileage
  - Fuel Type
  - Tax
  - Miles per Gallon
  - Engine Size

SUCCESS CRITERIA/METRICS

- Accuracy – root mean squared error (RMSE) below five thousand
  - Five thousand is ¼ of average used car price in historical dataset
- Efficiency – Can a seamless pipeline be created for baseline predictions?

ANALYSIS ASSUMPTIONS

- Five thousand RMSE = accuracy threshold for success
- Selling price = price analyst produced using available data
- All sales happened within short time interval
- Sales in data are representative of all BMW used car sales
- Prices do not have human error

ANALYSIS

- Accuracy – Model Development
  - What is it?
  - Initial Results
  - Final Results
  - Uncertainty
- Efficiency – Package Development
  - What is it?
  - Prediction Efficiency
  - Other Efficiency Perks

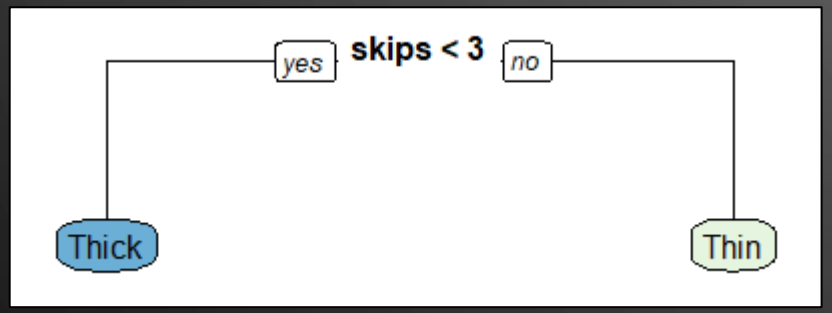
MODEL DEVELOPMENT – WHAT IS IT?

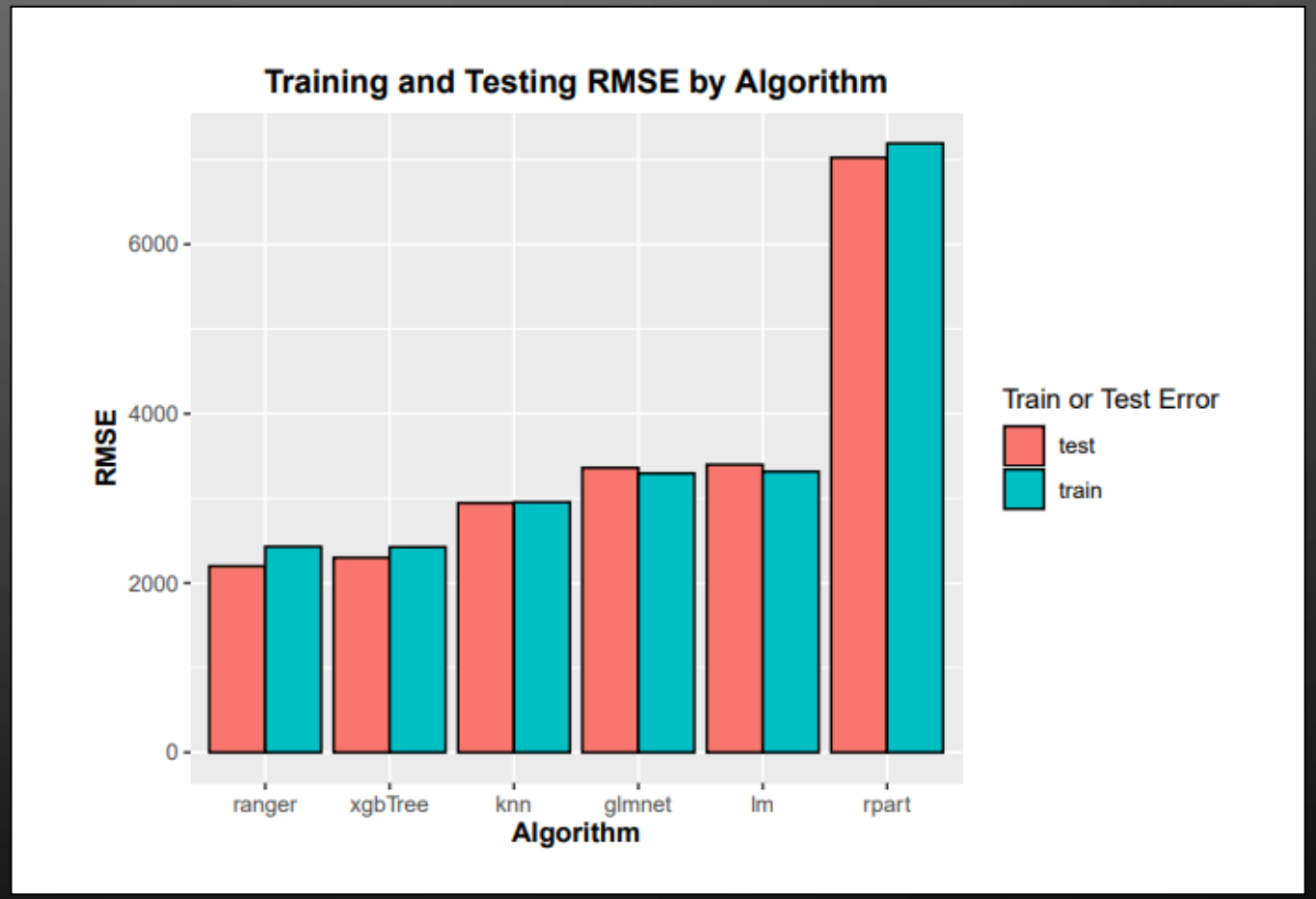
- Model = a set of rules learned from historical data to predict a baseline price
  - Like a pricing analyst using intuition to price a vehicle
- Different algorithms use data differently just like different analysts would

MODEL DEVELOPMENT – INITIAL ALGORITHM RESULTS

- Lower error the better – ranger algorithm performed best
- "ranger" = Random Forest

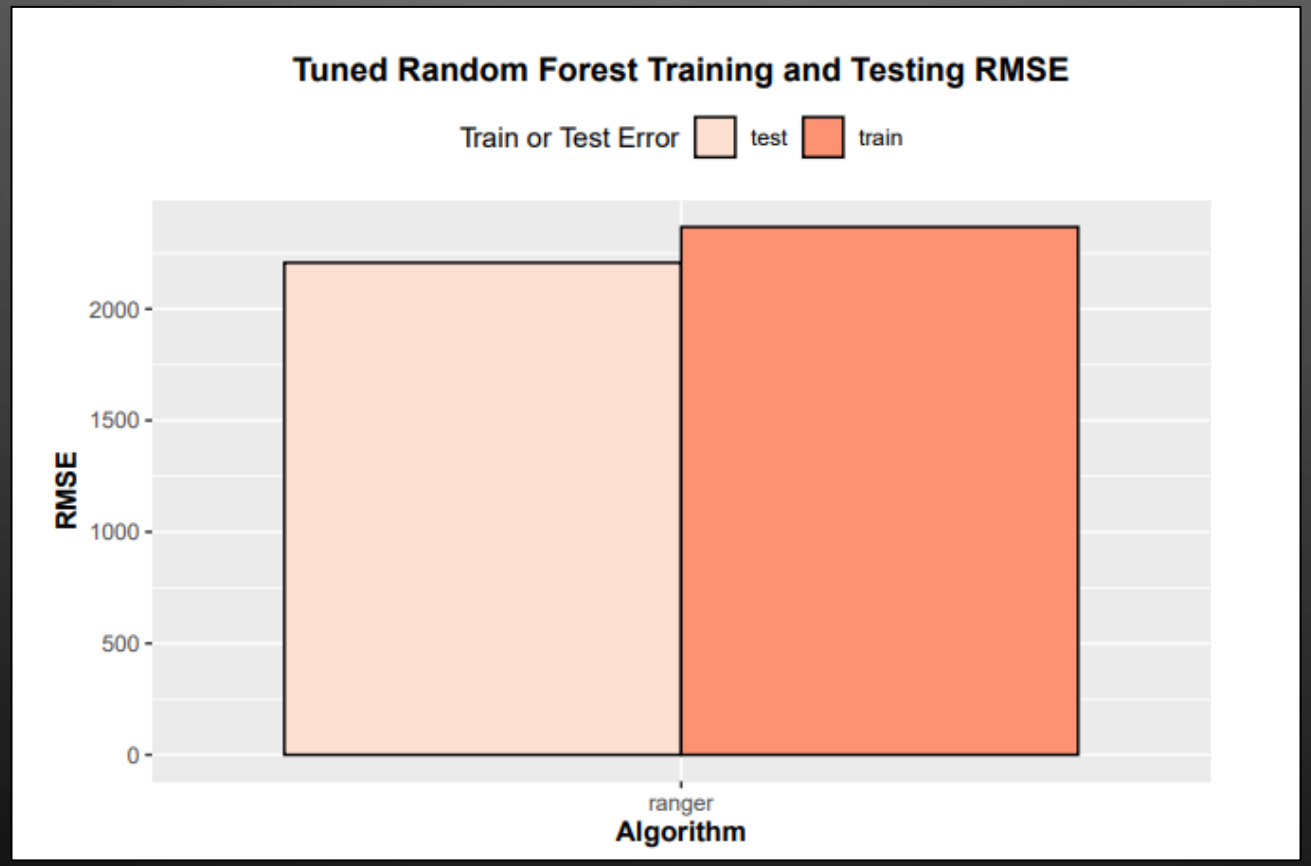
Example Decision Tree





MODEL DEVELOPMENT – FINAL RESULTS

- RMSE below five thousand, our metric for success



MODEL DEVELOPMENT – UNCERTAINTY

- RMSEs shown are *averages* over many observations
  - Not guaranteeing every prediction will have that error
  - Some predictions have larger errors while others have smaller ones

- 3,000 RMSE = On *average*, model is off by 3,000 units

PACKAGE DEVELOPMENT – WHAT IS IT?

- Means to ensure the same process is efficiently applied to incoming data for baseline price prediction

PACKAGE DEVELOPMENT – PREDICTION EFFICIENCY

- Individual prediction steps (all requiring different functions):
  - Feature generation
  - Dummy variable creation
  - Centering/Scaling
  - Model load
  - Price Prediction
- Final package results – two seamless steps (two functions)
  - One function prepares data for prediction
  - Second function predicts price

PACKAGE DEVELOPMENT – OTHER EFFICIENCY PERKS

- Deployable – GitHub, Azure DevOps
  - No individual sharing via email, Microsoft Teams, etc.
- Documentation
  - Saves questions/answers via email & meetings
- Version Control – Leverage Git
  - Cleaner than multiple Excel files in same directory
  - Easier to revert changes

RESULTS - SUCCESSFUL

- Can data science be used to not only generate a more accurate baseline price but also speed up the pricing process to better ensure company profitability?
  - Accuracy – Model able to generate baseline predictions well below the five thousand threshold
  - Efficiency – Package allows for seamless baseline prediction generation

NEXT STEPS/RECOMMENDATIONS

- Is five thousand the “correct” RMSE?
- Will structure of future data flowing through pipeline be the same?
- Will volume of future data be more/less?
- What other algorithms could be tested for modeling?
- How can the package be expanded to better suite the production environment?

QUESTIONS?