Industry Project: Wait Room Analysis

By: The Innovative Thinkers Group INC.
(Marika Fox, Najahme Ridley, Khalid Aboalayon, & Jonah White)

Introduction

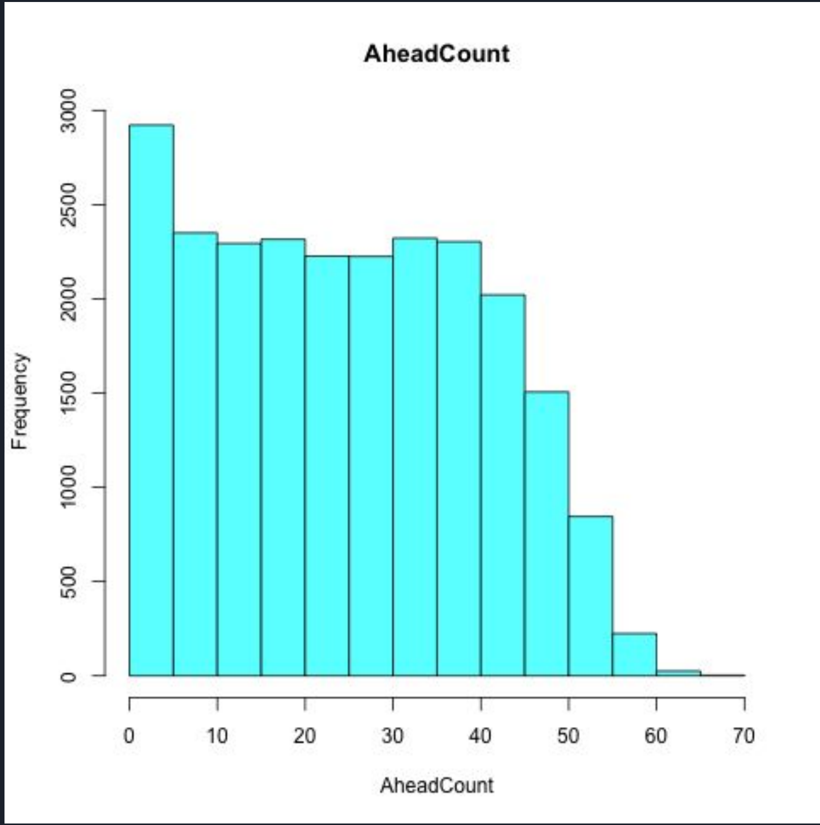
- Objective: To determine the cause of patient wait-times using data analysis and create a detailed model that predicts patient wait-time.

- For our standardized error metric we use Mean Squared Error (MSE).
  - MSE measures the actual squared difference between the estimated and actual values.
  - We will be using this metric to rate how well our models predicted the wait times.
- Identifying stakeholders: Medical facility staff, Executive team members, and incoming patients

Data Exploration
Histograms

- Reduce number of variables by assessing relevance of each
- Histograms to inspect distribution
- Histograms and qualitative analysis to eliminate 49 variables

Variables
84 → 35



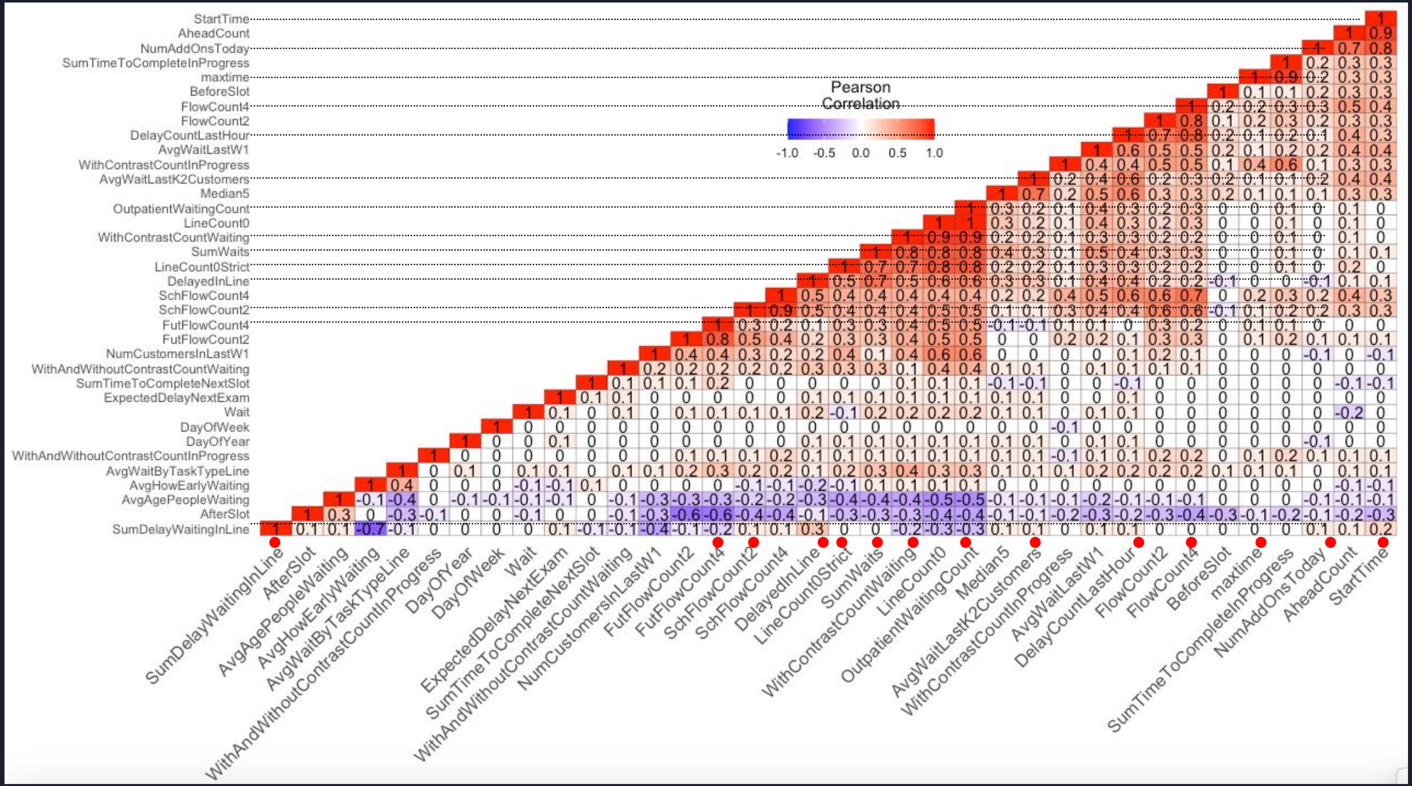
Data Exploration

Heat Mapping

- Eliminate redundant variables with |correlation| >0.7

Variables
35 → 21

*DayOfWeek
*DayOfYear
removed as well because they are not useful for regression analyses



Building The Model

Multiple Linear Regression (MLR)

- After we removed the highly correlated variables, we have 20 variables.
- Thus, we built our model for estimating Wait time (dependent variable) based on 19 predictors (independent variables) to answer the following:
  1. Does our set of predictors do a good job in predicting our outcome (wait time)?
  2. Which variables in particular are significant predictors of the outcome?

Model Summary

```
Call:
lm(formula = Wait ~ ., data = waitData)

Residuals:
    Min      1Q  Median      3Q     Max
-50.112 -18.317  -5.419  11.936 304.632

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     32.902722   3.643090   9.032  < 2e-16 ***
AvgHowEarlyWaiting              -0.043367   0.009282  -4.672 3.01e-06 ***
LineCount0                       1.265073   0.140301   9.017  < 2e-16 ***
FlowCount2                      -0.931897   0.218648  -4.262 2.04e-05 ***
SchFlowCount4                    1.248960   0.128452   9.723  < 2e-16 ***
FutFlowCount2                   -0.813847   0.243149  -3.347 0.000819 ***
AheadCount                      -0.360074   0.017547 -20.520  < 2e-16 ***
BeforeSlot                       0.066351   0.032776   2.024 0.042948 *
AfterSlot                       -0.154545   0.029978  -5.155 2.56e-07 ***
Median5                          0.114483   0.010390  11.019  < 2e-16 ***
AvgWaitByTaskTypeLine            0.016543   0.018144   0.912 0.361915
SumTimeToCompleteInProgress     -0.004327   0.004359  -0.993 0.320927
ExpectedDelayNextExam            0.132533   0.051493   2.574 0.010067 *
AvgAgePeopleWaiting             -0.043255   0.037685  -1.148 0.251063
NumCustomersInLastW1            -0.669490   0.169573  -3.948 7.91e-05 ***
AvgWaitLastW1                    0.054220   0.010426   5.201 2.01e-07 ***
SumTimeToCompleteNextSlot        0.031395   0.021283   1.475 0.140199
WithAndWithoutContrastCountWaiting 0.353499 0.302063   1.170 0.241905
WithContrastCountInProgress     -0.187271   0.363251  -0.516 0.606181
WithAndWithoutContrastCountInProgress -0.350188 0.560722 -0.625 0.532288
```

Interpreting the MLR analysis

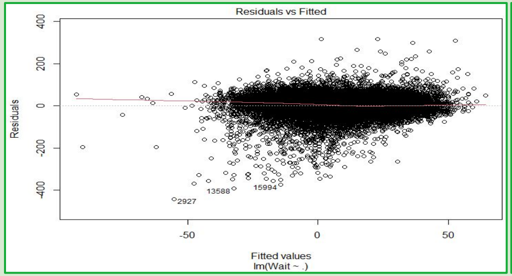
The p-value of the F-statistic is < 2.2e-16, which is highly significant

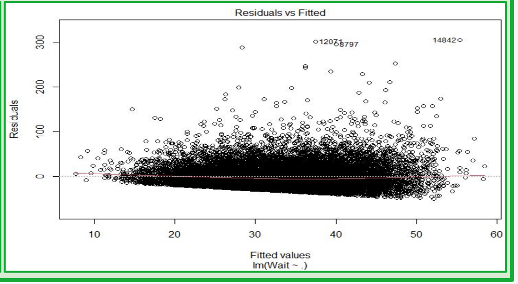
Meaning: at least one of the predictor variables is significantly related to the outcome (Wait time)

Residual Standard Error

with neg. values = 45.03

without neg. values = 26.99





Applying Models

Data Preparation:

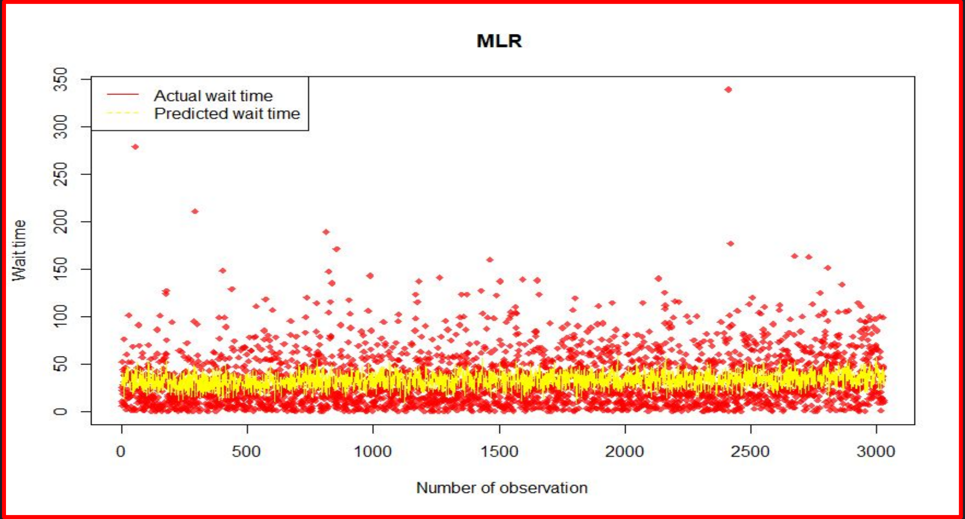
The datasets split randomly with train data containing 80% of the data and 20% for testing data.

Three regression models are used:

1-Multiple Linear Regression.
2-Support Vector Regression
3-Random Forest Regression

First: Multiple Linear Regression (MLR) Modeling

Evaluate MLR	
MAE	19.76956
MSE	705.1798
RMSE	26.55522

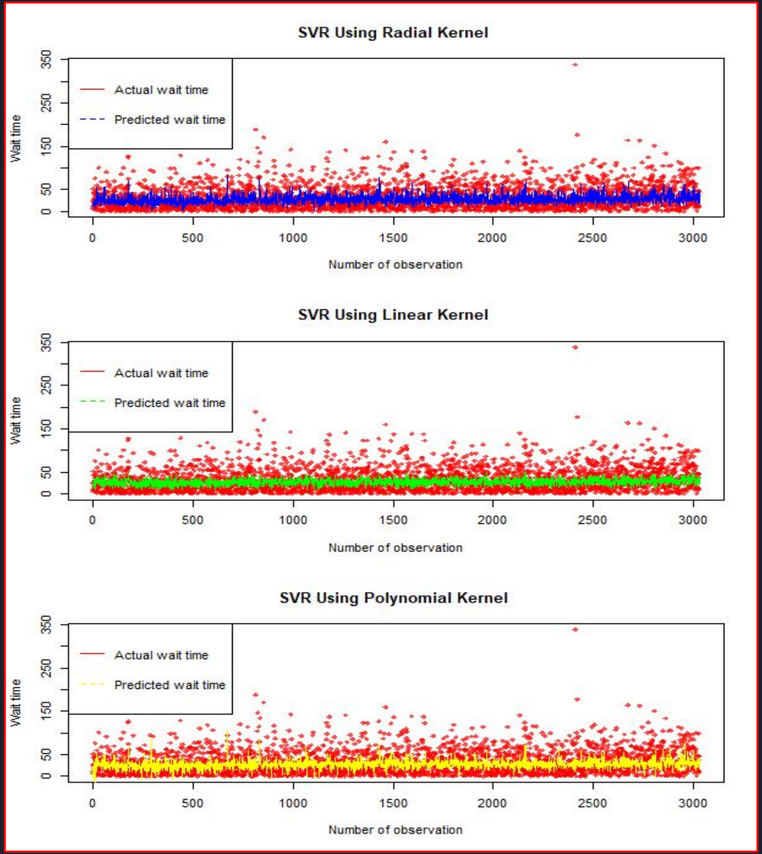


Second: Support Vector Regression (SVR)

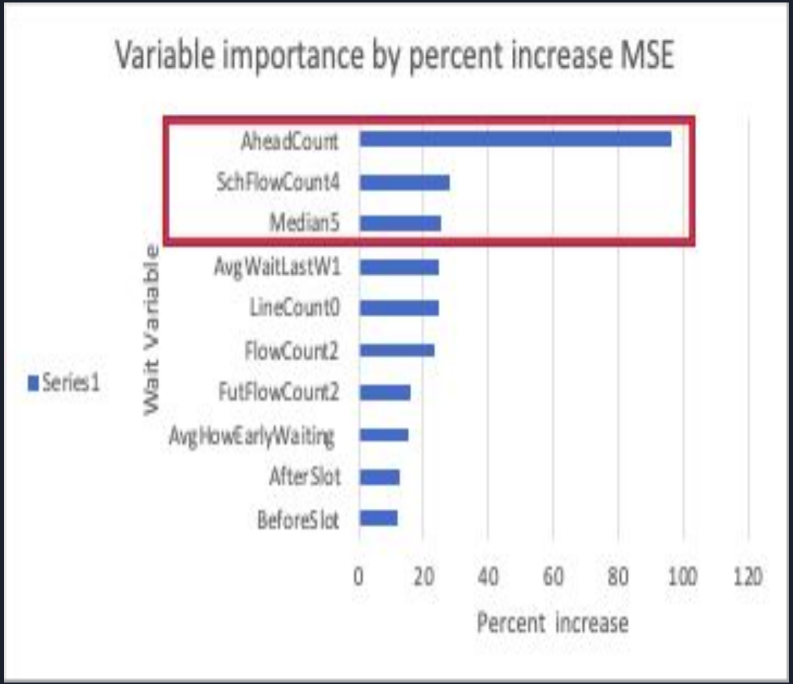
How to Build a Support Vector Regression Model:

1. Collect a training set.
2. Choose a kernel
3. Train the model to get contraction coefficient.
4. Use this coefficient to create an estimator/predictor.

SVM - Kernel	Metrics		
	MAE	MSE	RMSE
Radial	19.4195	668.3978	25.85339
Linear	19.04545	729.1731	27.0032
Polynomial	19.08464	710.6518	26.65805

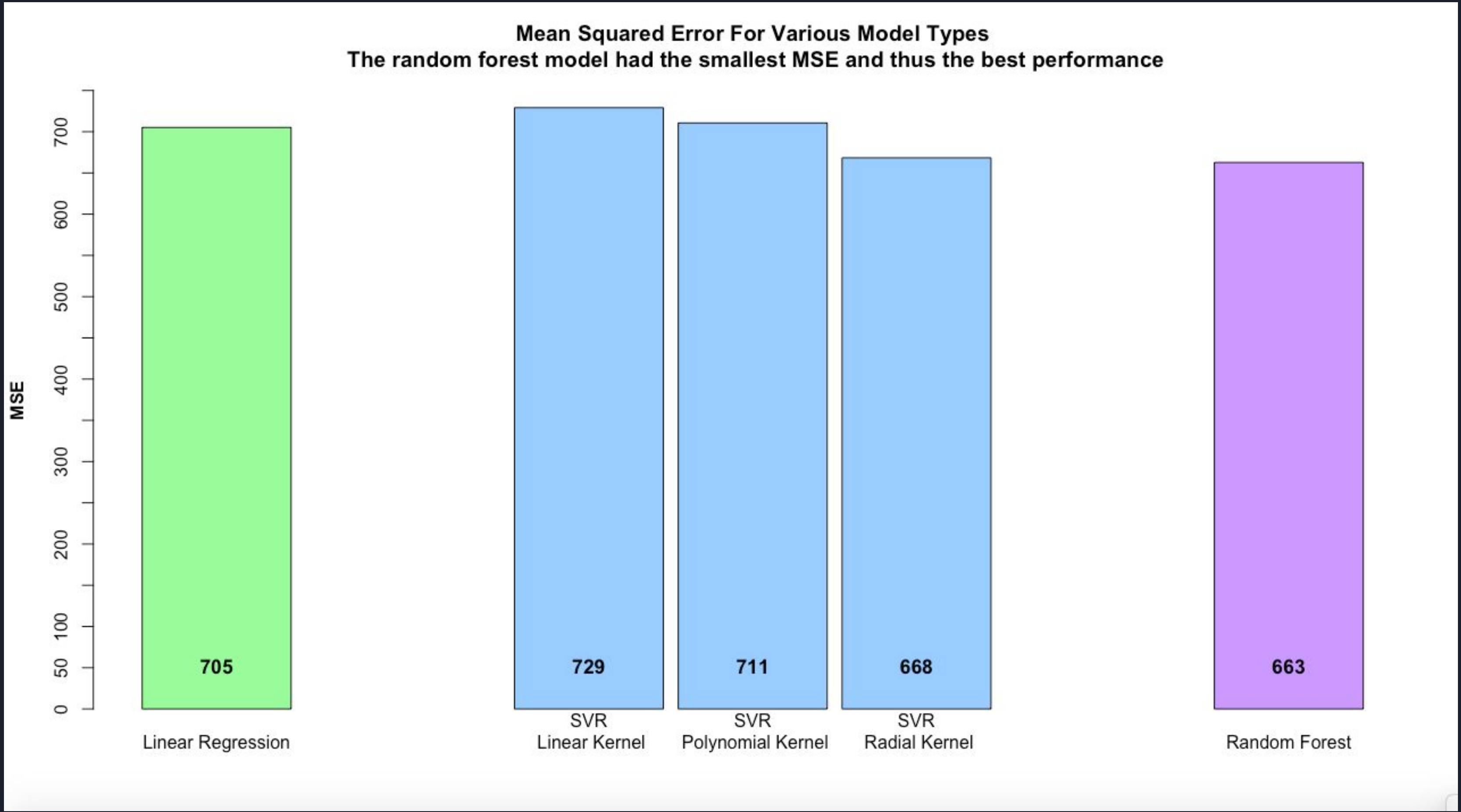


Random Forest Modeling



- We use Random Forest to determine which variables were most heavily weighted in our model.
- In staying consistent with the previous analysis., we removed variables with high p-values from our calculation.
- For this model additional time was taken to tune the parameters to optimize the MSE.
- Random Forest was our most successful model with an MSE of 663..
- We we able to identify the 3 most important variables to our model.
    - Ahead count
    - SchFlowCount4
    - Median5

Model Performances Summary



Observations

- Most successful model: Random Forest
  - Smallest MSE
- SVR (Radial Kernel)
- Linear Regression

Most Important Variables

1. Number of patients scheduled before current patient for the day.
2. Number of patients scheduled in the 60-minute window before patient arrived.
3. Median delay/wait time for 5 most recent customers.

Recommendations

- All of the most significant variables relate to patient traffic, not hospital resources
  - Initiatives to reduce wait time must focus on improving scheduling and movement of patients
- Our suggestions:
  - Track causes for delayed wait times i.e. paperwork, proof of insurance, etc. These variables can help us make procedural recommendations.
  - Track the medians of more variables in addition to averages, to help track more variations over time
- Next steps for model improvement:
  - Separate outliers that deviate by +30 minutes from the median data. These are extenuating circumstances that should not typically occur.

Resources

Dataset: https://medicalanalytics.group/operational-data-challenge/

Heat-map explanation: https://stats.stackexchange.com/questions/392517/how-can-one-interpret-a-heat-map-plot:

Random Forest explanation:https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/tutorial-random-forest-parameter-tuning-r/tutorial/

P-value Explanation:

https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients#:~:text=How%20Do%20I%20Interpret%20the,can%20reject%20the%20null%20hypothesis.

Google Images

SVM - Kernel	High p-Value Variables	Metrics		
		MAE	MSE	RMSE
Radial	with	19.4195	668.3978	25.85339
	without	19.17606	713.1397	26.70468
Linear	with	19.04545	729.1731	27.0032
	without	19.06767	731.3685	27.04383
Polynomial	with	19.08464	710.6518	26.65805
	without	19.69521	756.1212	27.4976