**Final Project 10 Team (Bamboo Consulting): Julia So, Elena Xiao, Zoey Zhao, Carol Zhou**

**Data Set**

We will be using the 2019 data set, “Airlines Customer Satisfaction” ( <https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction> ). As people begin to travel more post pandemic, airlines are looking to re-capture and retain customers through a positive experience. This test tracks different aspects of a customer’s experience and whether the customer was satisfied or not.

There are 129,881 instances in the data set. In total, there are 23 features. The target variable is *satisfaction* and is categoric*.* The other 22 are a mix of categorical (4) and numeric (18) features. The only variable with missing instances is *Arrival.Delay.in.Minutes* with 393 missing instances. Those values will be imputed with the median.

**Who’s Who**

For privacy reasons, the actual airline has been referred to as Invisitico. We will continue to use this name. In this project, we are data analysts from Bamboo Consulting Firm hired by Invisitico’s Customer Insights and Experience team.

**Problem to Solve**

The business problem we are trying to solve is how to predict whether or not a customer will be satisfied with their airline experience. Understanding which features predict satisfaction can help the airline identify how they can improve their customer’s experience or intervene if it seems like a customer might not be satisfied. In data mining terms, this is a binary classification problem (classifying satisfied vs. dissatisfied customers).

**Planning Ahead**

**\*All use seed 42**

The models we plan to test are logistic regression-Kefeng, decision tree-Zoey, random forest-Carol, SVM (radial basis, polynomial)-Julia, and ANN (b/c target variable is binary categoric)-Kefeng.

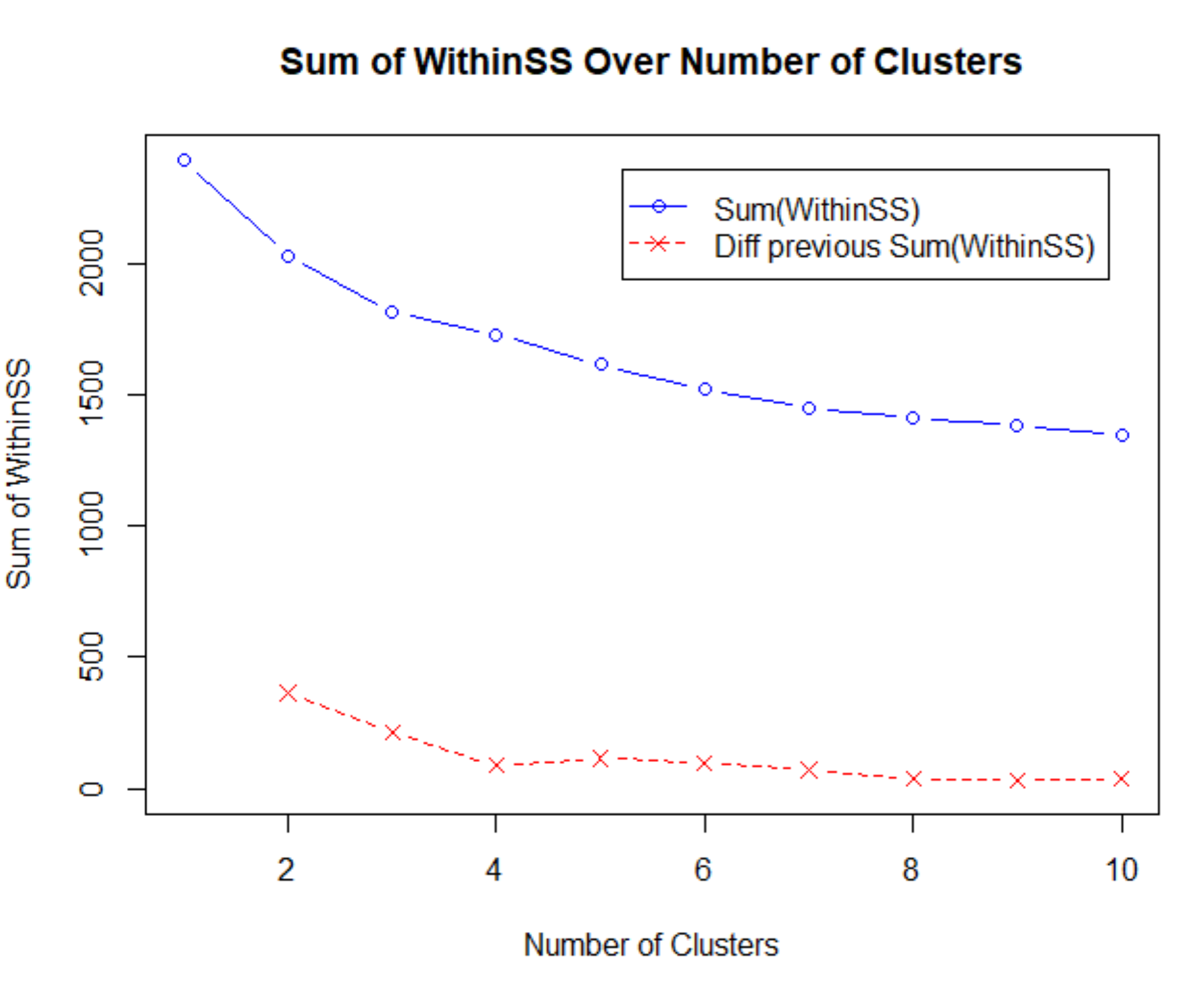
Gradient boost- Zoey

Adaboost- Julia

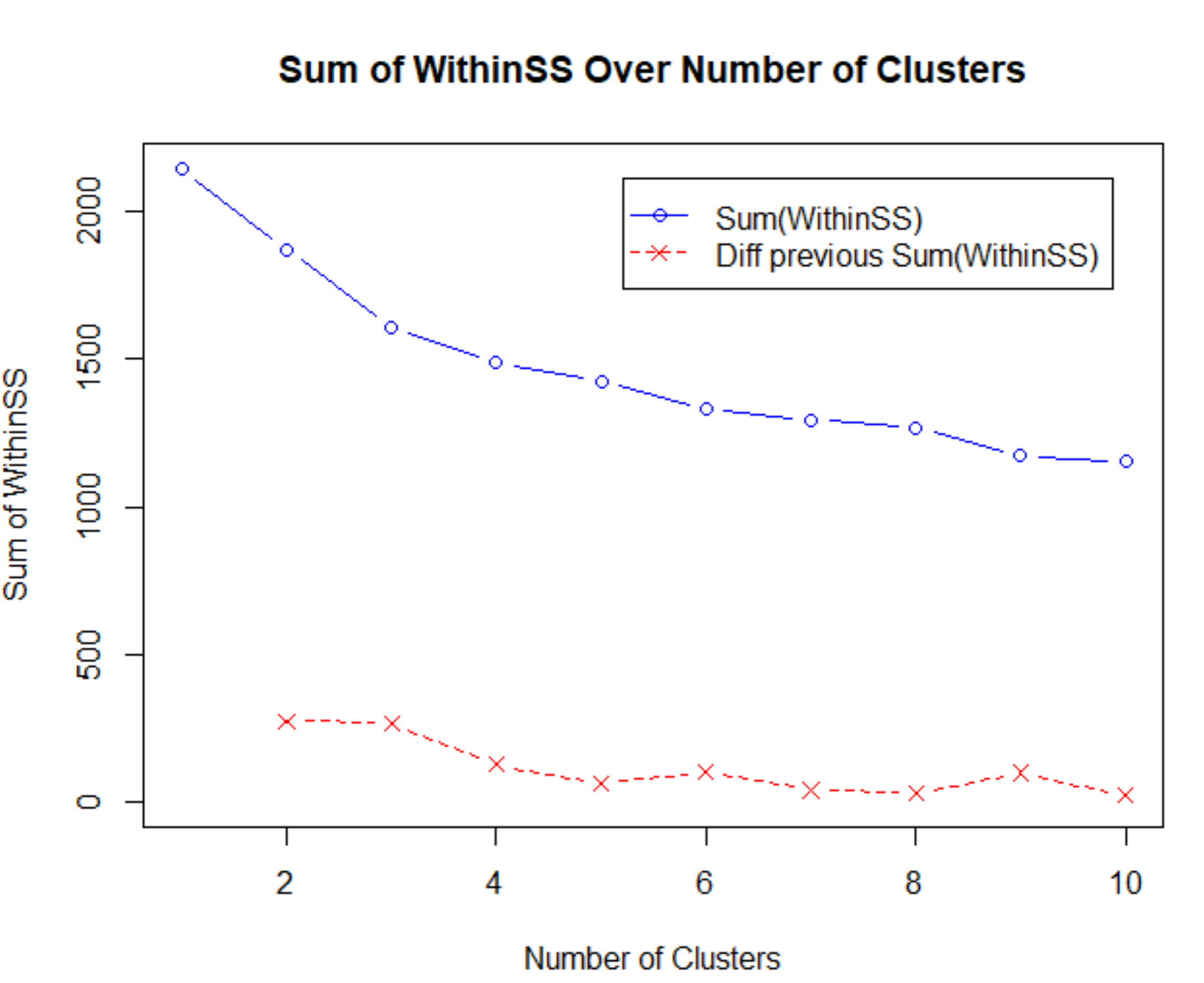
We will be using the validation set to determine which model has the highest AUC. The model with the highest AUC is the one we will recommend. The best model will then be ran with the test set and its AUC will be provided.

**K-cluster- scaled numeric from 0-1, categoric to indicator - Julia**

Dissatisfied

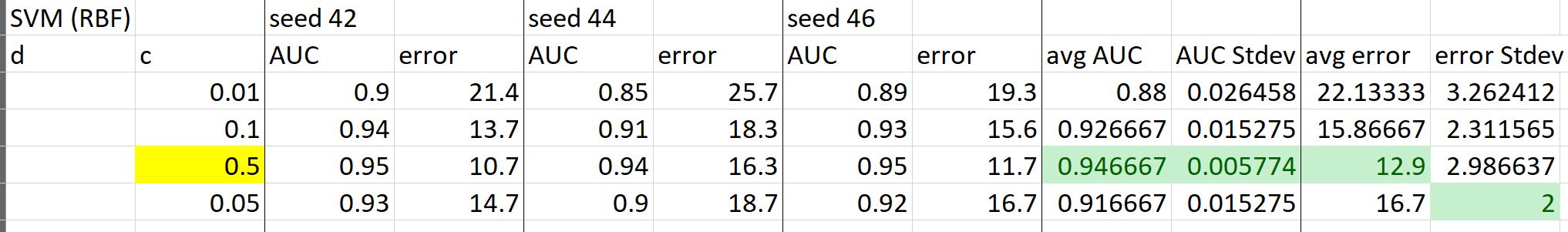


Satisfied- Julia



3 Clusters

**SVM- all features**



SVM (Polynomial)

|  |  |  |
| --- | --- | --- |
| Degree | Complexity | Validation AUC |
| 1 | .01 | .9198 |
| 1 | .1 | .9179 |
| 1 | .5 | .9178 |
| 2 | .01 | .9527 |
| 2 | .1 | .9400 |
| 2 | .5 | .9303 |

SVM (RBF)

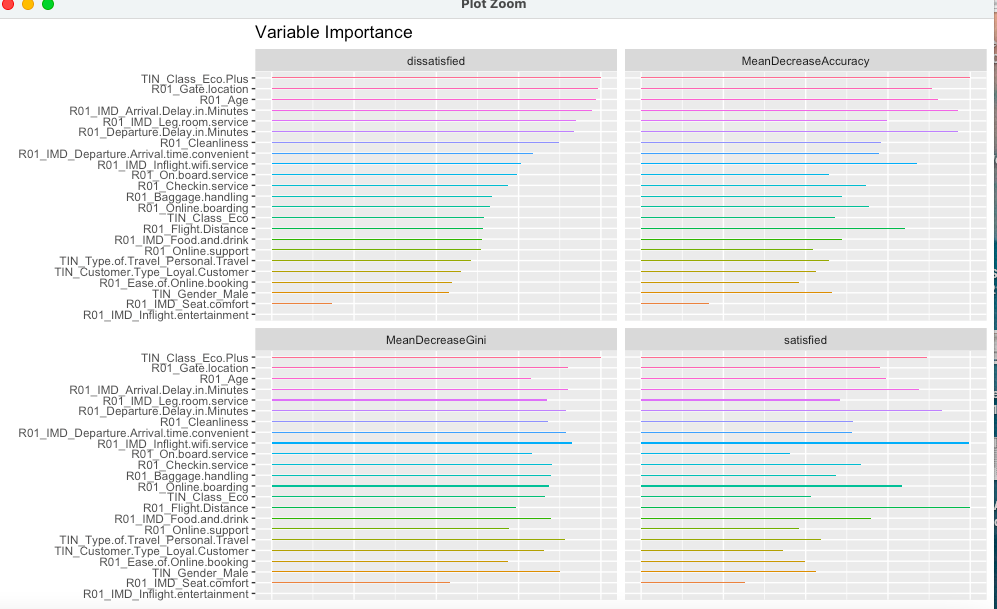
|  |  |
| --- | --- |
| Complexity | Validation AUC |
| .01 | .8994 |
| .05 | .9261 |
| .1 | .9363 |
| .25 | .9467 |
| .5 | .9520 |

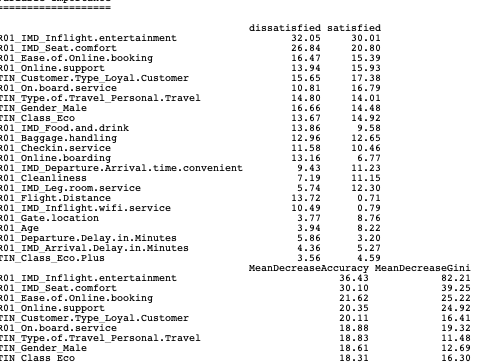
**Random Forest- all features**

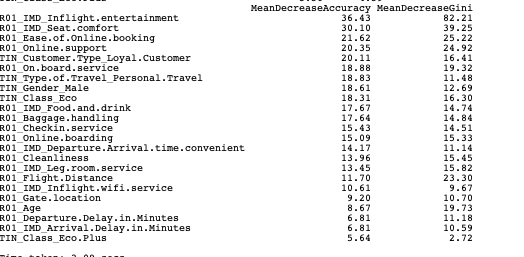
**Test: 0.9690**

|  |  |  |
| --- | --- | --- |
| Tree | Variable | Validation AUC |
| 500 | 4 | 0.9592 |
| 500 | 5 | 0.9602 |
| 500 | 6 | 0.9577 |
| 500 | 3 | 0.9607 |
| 400 | 3 | 0.9575 |
| 300 | 3 | 0.9583 |

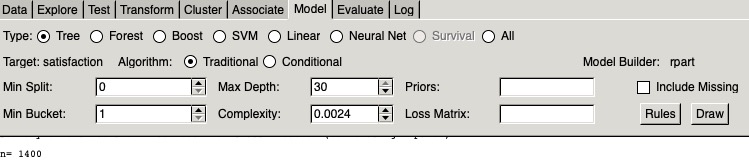
Tt

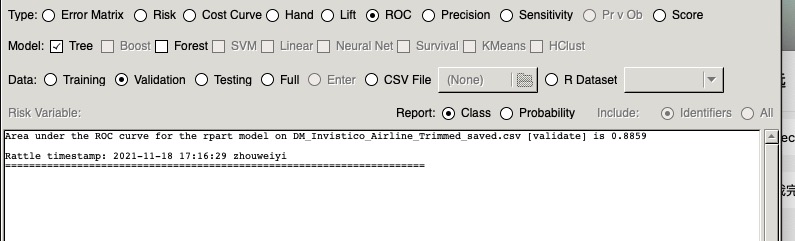




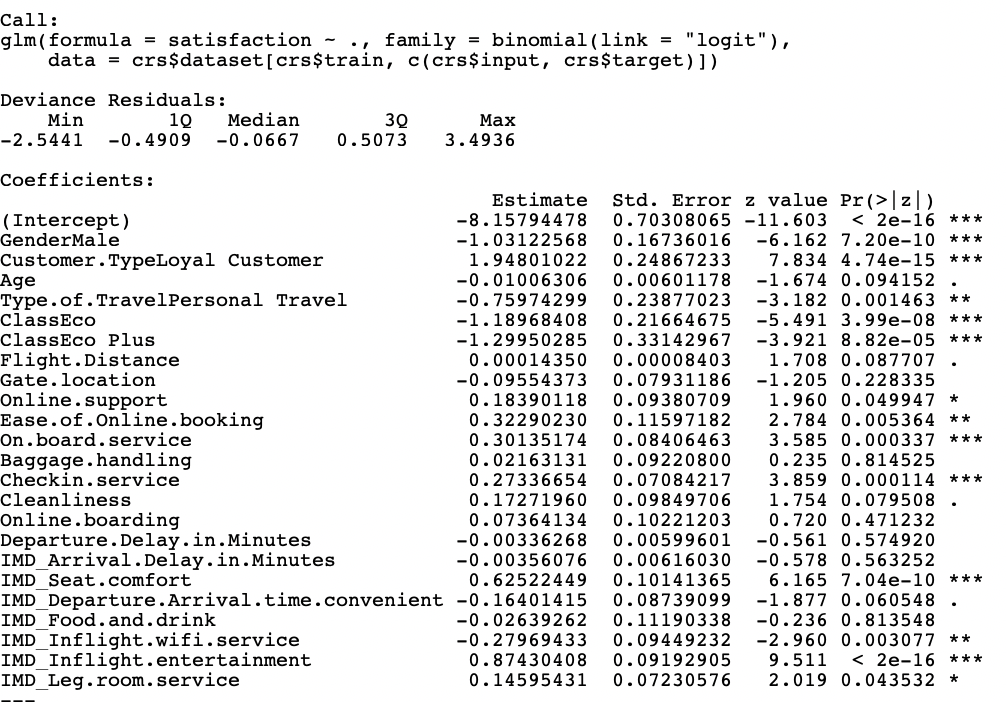


**Decision Tree features**



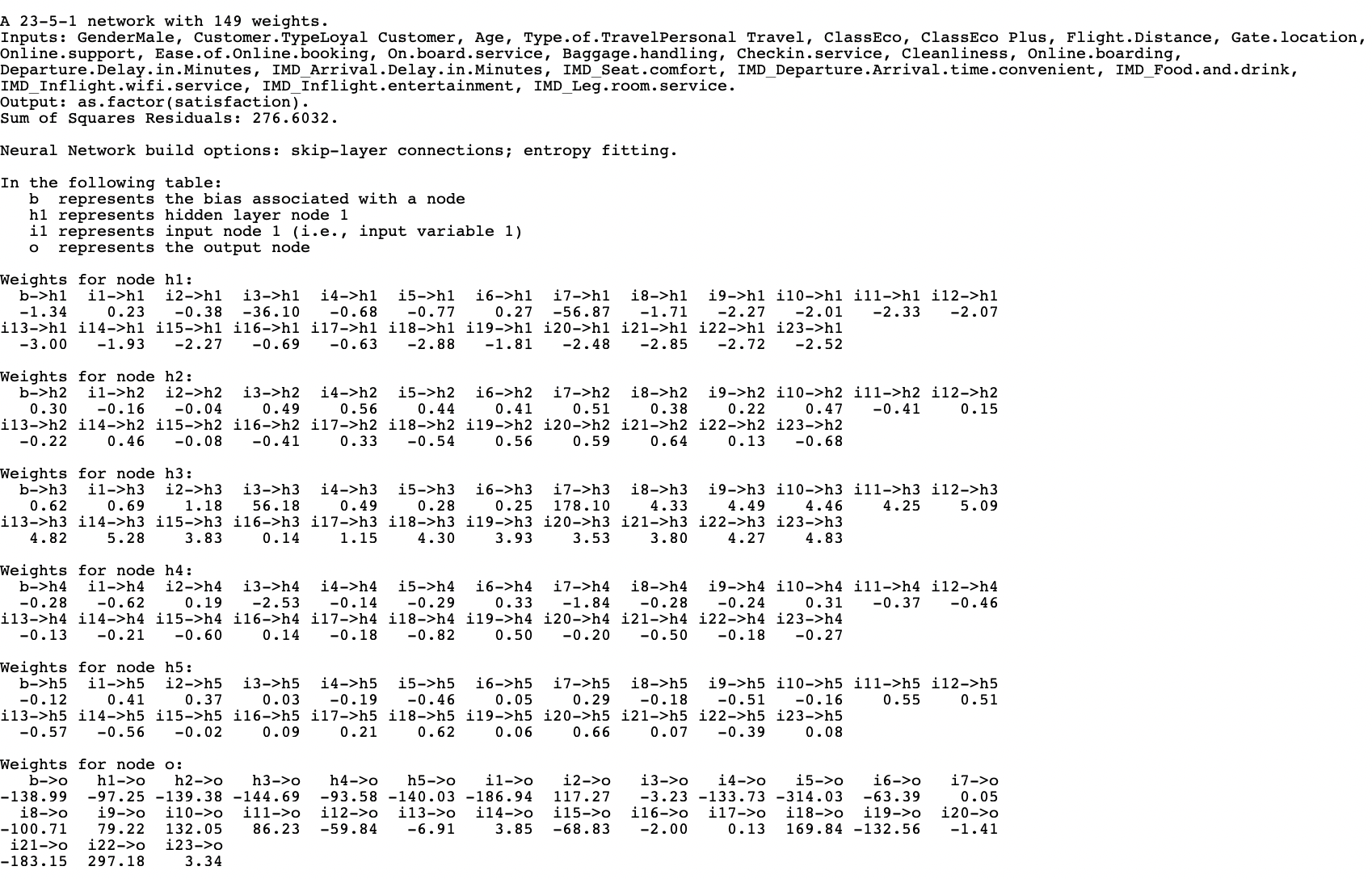


**Logistic regression**



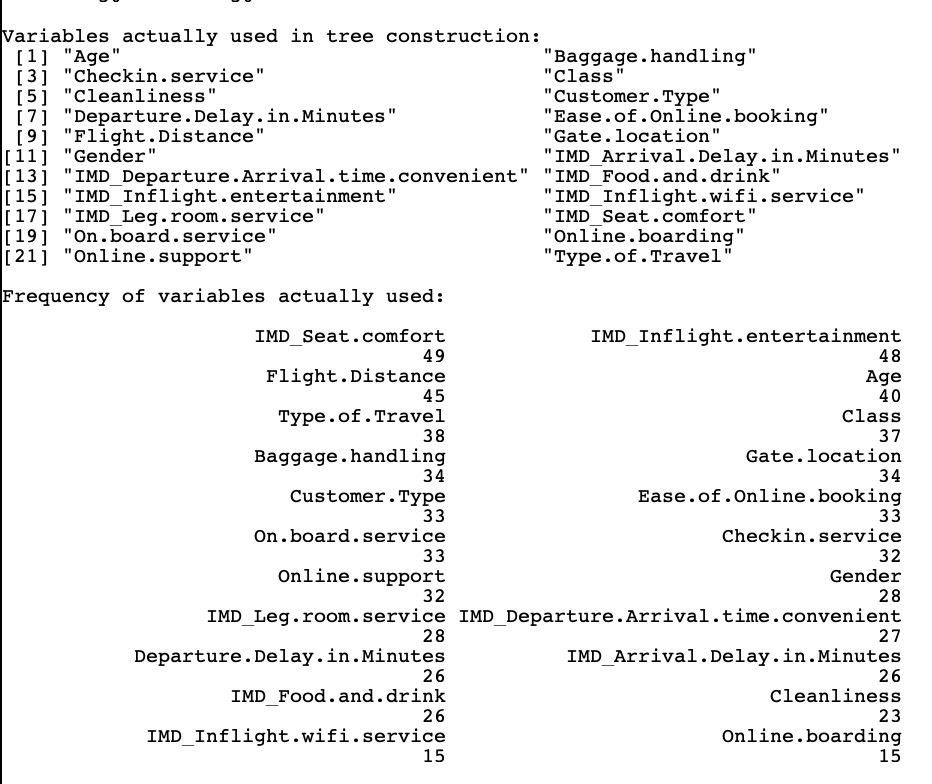


ANN

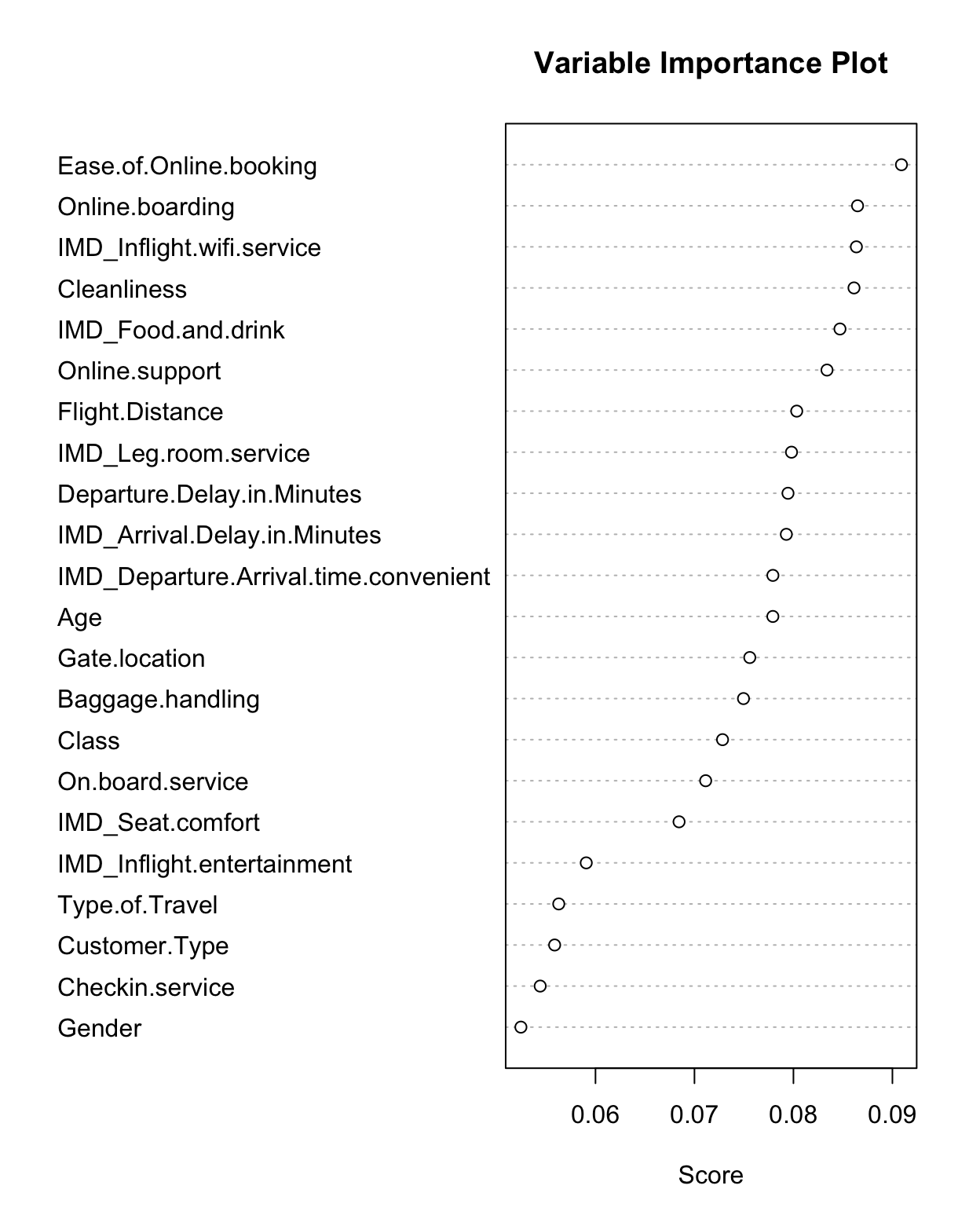




Ada Boost







Gradient Boost

