LINK TO REPOSITORY: https://github.com/rosenshabgu/ex2

Algorithms used for this ensemble: xgbTree and C5

Pre-processing:

* Removed unintelligible features such as Name, Ticket, Cabin and PassengerId from Test and Train datasets
* We decided Cabin was unnecessary for the model because it had more than 70% missing values.
* Converted PClass from numeric feature to categorical for Train and Test dataframes
* Replaced categorical columns with binary "dummy" columns, such as PClass, Sex and Embarked
* From the Train dataframe, we omitted all rows that still had Na values

Algorithms used:

* xgbTree
* c5.0

Parameter selection:

Initialized a seed to a value of 1

Metric of choice: Accuracy (for both algorithms)

For c5.0:

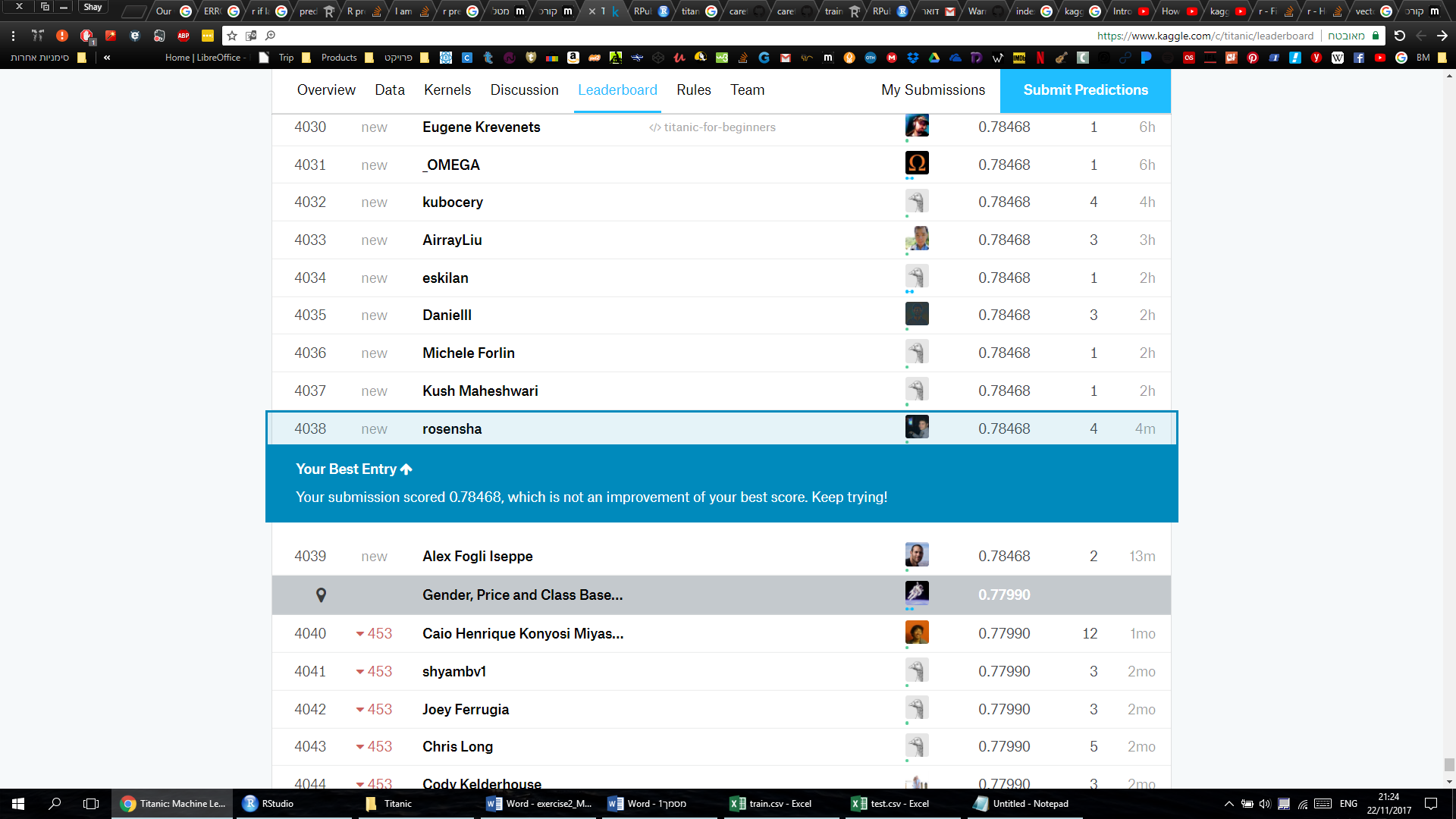
* Winnow set to TRUE
* Trials set to 2

For xgbTree:

* Nrounds set to 20
* Max\_depth set to 7
* Eta set to 0.3
* Gamma set to 0.1
* Colsample\_bytree set to 0.7
* Min\_child\_weight set to 0.01
* Subsample set to 0.6

Link to github repository:

Print screen image:



Basic algorithm #1 – C 5.0

Pre-processing:

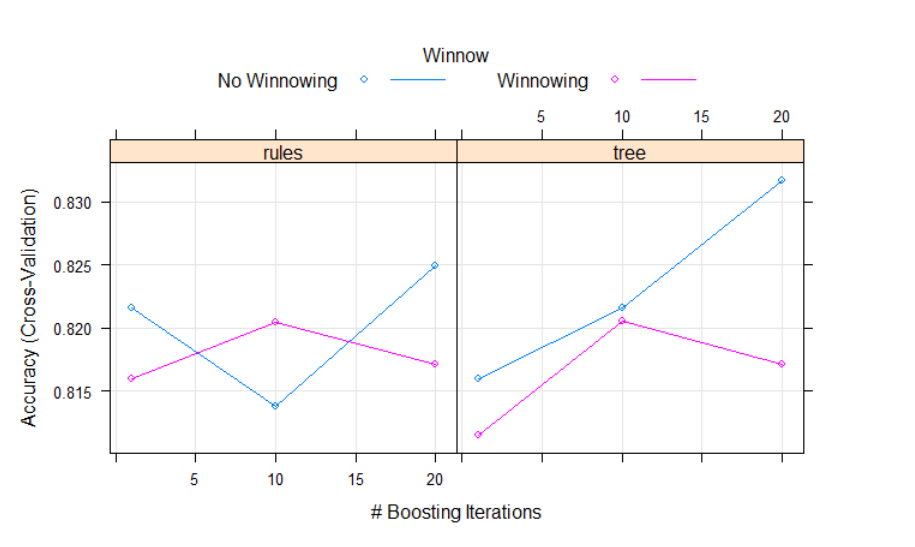
* Removed unintelligible features such as Name, Ticket and PassengerId from Test and Train datasets
* Converted PClass from numeric feature to categorical for Train and Test dataframes
* Converted Survived to categorical on the Train set
* Replaced categorical columns with binary "dummy" columns, such as PClass, Sex and Embarked
* From the Train dataframe, we omitted all rows that still had Na values

Algorithms used:

* xgbTree
* c5.0

Parameter selection:

* Seed set to 1
* Metric of choice is Accuracy
* Winnowing set to (True, False)



Print screen image:



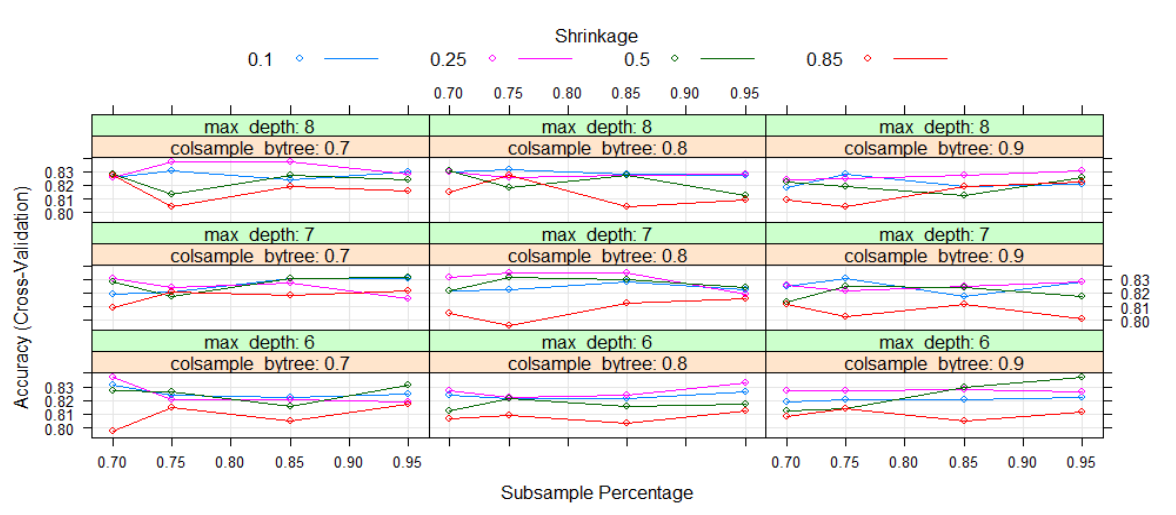
Basic algorithm #1 – XGBoost Tree

Pre-processing:

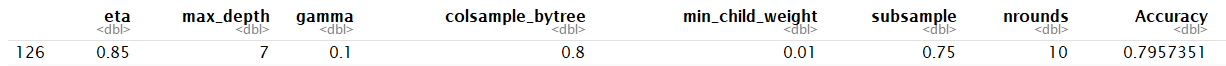
* Removed unintelligible features such as Name, Ticket, Cabin and PassengerId from Test and Train datasets
* We decided Cabin was unnecessary for the model because it had more than 70% missing values.
* Converted PClass from numeric feature to categorical for Train and Test dataframes
* Replaced categorical columns with binary "dummy" columns, such as PClass, Sex and Embarked

Parameter selection:

* Seed set to 123
* Metric of choice is Accuracy
* Nrounds set to 10
* Max\_depth set to (6, 7, 8)
* Eta set to (0.1,0.25,0.5,0.85)
* Colsample\_bytree set to (0.8,0.9,0.7)
* Min\_child\_weight set to 0.01
* Subsample set to (0.75, 0.85, 0.95, 0.7)
* 3 fold Cross Validation
* Na handling: Pass



The best Accuracy was given with the following parameters:



Print screen image:

C:\Users\dell\AppData\Local\Microsoft\Windows\INetCache\Content.Word\xgbscore.png