

EE627-WS Data Acquisition and  
Processing I Final Project Kaggle Remark

**Team Name: Star Wars**

**Justin John, Ismet Okan Celik,**

**Anthony Donatelli**

Prof. Wang

05/13/21

# Kaggle Submissions:

If the project had errors the team resubmitted accordingly. These are the list of Kaggle Submissions which brief remarks about which methods were used for said files and so on.

Format:

**File name** (*Public Score*): Remark

**FEATURE\_ENGINEERING.csv** (0.83941)

This csv file has the following: 8 features, using mean, std, median, and other statical values. Find the best track for each user.

**TAKING\_MAX\_RATES.csv** (0.57959)

Applying Max Function and taking highest 3 and change with 1 rest 0. The professor recommended we try this approach and this is the resulting csv.

**PCA\_RESTRUCTURING\_PREDICTION.csv** (0.84313)

The original values were submitted with values ranging from 0-100 and then they were converted. The team took the best three tracks for each user and this was the file that was produced.

**OUTPUT\_15.csv** (0.47281)

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**pca weighted output 7.csv** (0.86662)

Applied a global threshold of 10 and a per user threshold. This was used to adjust our original data that ranged from 0-100 to values containing only 0s and 1s.

**pca weighted output 6.csv** (0.86813)

Applied a global threshold of 15 and a per user threshold. This was used to adjust our original data that ranged from 0-100 to values containing only 0s and 1s.

**pca weighted output 5.csv** (0.86153)

Applied a global threshold of 20 and a per user threshold. This was used to adjust our original data that ranged from 0-100 to values containing only 0s and 1s.

**Ensemble\_Predictions\_Test\_11\_values.csv** (0.85563)

The team used the best 11 submissions they had and used the assembling method taught in class to produce this file.

**pca weighted output 4.csv (0.86679)**

Applied a global threshold of 25 and a per user threshold. This was used to adjust our original data that ranged from 0-100 to values containing only 0s and 1s.

**Ensemble\_Predictions\_Test\_10\_values.csv (0.85217)**

The team used the best 10 submissions they had and used the assembling method taught in class to produce this file.

**pca weighted output 3.csv (0.86259)**

For each user, average predictor values are calculated and this value is set as a threshold and above the threshold assigned as 1 and the rest of the values assigned as 0's.

**pca weighted output 2.csv (0.74126)**

For each user, average predictor values are calculated and this value is set as a threshold and above the threshold assigned as 1 and the rest of the values assigned as 0's.

**PCA\_1.csv (0.77795)**

For each user, average predictor values are calculated and this value is set as a threshold and above the threshold assigned as 1 and the rest of the values assigned as 0's.

**PCA.csv (0.77795)**

For each user, average predictor values are calculated and this value is set as a threshold and above the threshold assigned as 1 and the rest of the values assigned as 0's.

**Reverse\_PCA.csv (0.50307):**

We took one of our previous submissions and changed 1's to 0' and 0's to 1's.

**Ensemble\_Predictions\_Test5.csv (0.84501)**

The team used the best 5 submissions they had and used the assembling method taught in class to produce this file.

**Ensemble\_Predictions\_Test4.csv (0.84501)**

The team used the different 5 previous submissions they had and used the assembling method taught in class to produce this file.

**Feature\_Engineering\_updated\_1.csv (0.77505)**

Feature engineering applied and created new features and some of the better features added each other and submitted.

**PCA\_UPDATED2.csv (0.77795):**

One of our previous submissions is converted in to 1 and 0's and submitted

**Ensemble\_Predictions\_Test3.csv (0.50170)**

The team used the worst 3 submissions they had and used the assembling method taught in class to produce this file.

**Ensemble\_Predictions\_Test2.csv (0.50000)**

The team used the worst 4 submissions they had and used the assembling method taught in class to produce this file.

**OUTPUT\_14.csv (0.78478)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_13.csv (0.55624)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_12.csv (0.77707)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_11.csv (0.79575)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_10.csv (0.75773)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_9.csv (0.75291):**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**Ensembling\_Learning.csv (0.68595)**

The team used the 2 good submissions and 2 bad submissions they had and used the assembling method taught in class to produce this file.

**OUTPUT\_8.csv (0.75109)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_6.csv (0.75083)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_5.csv (0.80648)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_4.csv (0.75493)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_3.csv (0.83257)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_2.csv (0.76905)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**OUTPUT\_1.csv (0.84991)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this.

**Gradient\_boosted\_tree.csv (0.84398)**

Using this gradient boosted tree algorithm taught in class, this file was produced.

**Random\_forest\_1.csv (0.82108)**

Using the Random Forest algorithm taught in class, this file was produced.

**Decision\_tree.csv (0.82108)**

Using the Decision Tree Classifier algorithm in class, this file was produced.

**New\_sub1.csv (0.77015)**

After talking to the professor about our score, we used a threshold to adjust the high values.

**New\_sub1.csv (0.74126)**

After talking to the professor about our score, we used a threshold to adjust the high values.

**Logical\_regression.csv (0.84507)**

Using the Logistic regression algorithm in class, this file was produced.

**Pyspark\_Summation\_All\_Features.csv (0.50804)**

Using the summation of All Features and Matrix Factorization to make this file.

**Matrix\_factorization\_attempt1.csv (0.49474)**

This was the team's first attempt at matrix factorization.

**OUTPUT\_SummaionFeatures\_Feature\_Engineering.csv (0.89210)**

Feature engineering applied and created new features (approx. 21 features) and all features added each other and submitted.

**OUTPUT\_Decomposition\_Feature\_Engineering.csv (0.61561)**

Feature engineering applied and created new features and some of the better features added each other and submitted.

**OUTPUT\_PCA\_Feature\_Engineering.csv (0.61561)**

Used Feature Engineering with a combination of PCA to make said file.

**pca weighted output.csv (0.89658)**

Using pca weights for an average of scores to produce the following result. This is one of the files that used values higher than 1s and 0s, so it was revisited and resubmitted as a new file.

**8features\_final\_results\_as\_id\_predictor\_values.csv (0.86602)**

This was the team's first attempt at using the 8 feature and seeing what type of accuracy score the team could get.

**OUTPUT5.csv (0.87205)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this. Including Genre1 and Genre2.

**OUTPUT3.csv (0.88247)**

This is one of the original submissions that was converted to 1s and 0s. Originally, the data was in values ranging from 0-100 and the professor told us to fix this. Including Genre1 to the calculation.

**0.95\_0.15\_output14.csv (0.87871)**

The team played around rearranging weighted average calculation to see what scores would be produced for said file. This was a weighted average file.

**0.90\_0.10\_output13.csv (0.87873)**

The team played around rearranging weighted average calculation to see what scores would be produced for said file. This was a weighted average file.

**0.85\_0.15\_output12.csv (0.87871)**

The team played around rearranging weighted average calculation to see what scores would be produced for said file. This was a weighted average file.

**0.80\_0.20\_output11.csv (0.87869)**

The team played around rearranging weighted average calculation to see what scores would be produced for said file. This was a weighted average file.

**0.75\_0.25\_output10.csv (0.87866)**

The team played around rearranging weighted average calculation to see what scores would be produced for said file. This was a weighted average file.

**0.3\_0.7\_output9.csv (0.87455)**

The team played around rearranging weighted average calculation to see what scores would be produced for said file. This was a weighted average file.

**Output8.csv (0.87863)**

This was our submission with taking weighted averages and finding predictor scores between 0 and 100. This was corrected to be 0's and 1's later.

**Output7.csv (0.84310)**

This was our submission with taking weighted averages and finding predictor scores between 0 and 100. This was corrected to be 0's and 1's later.

**Output5.csv (0.77585)**

This was our submission with taking weighted averages and finding predictor scores between 0 and 100. This was corrected to be 0's and 1's later.

**Output6.csv (0.77585)**

This was our submission with taking averages and finding predictor scores between 0 and 100. This was corrected to be 0's and 1's later.

**Output3.csv (0.87733)**

This was our first submission with taking simple averages and finding predictor scores between 0 and 100. This was corrected to be 0's and 1's later.