| Daisy away Oct. 17-26  Iris unavailable on Oct 22, 23, 29.  Amanda Oct. 18-24  Shuo unavailable on Oct 25, Nov. 11-13 | | |
| --- | --- | --- |
| **DELIVERABLE:** | **OUR TARGET DEADLINE** | **ACTUAL DEADLINE** |
| ~~Finalize groups & dataset~~ | ~~Oct. 5~~ | ~~Oct. 12~~ |
| ~~Present baseline submission~~  ~~(10 minute presentation)~~ | ~~Oct. 28~~ | ~~Nov. 2~~ |
| Present in class (15-20 minute presentation) | Dec. 2 | Dec. 7 |

## Meeting on 11/30

Agenda:

* Balance and then split or vice versa?
  + Spit and balance on the training.
  + Amit: There's no right or wrong answer. Split before or after is designed when we go train our model to combat over or under fit. People can argue different methods with under and over sampling. We’re doing the best we can with imbalanced data. A lot of times the dataset itself will suggest the answer. Keep it in mind and see what our counts are like.
  + Amanda: If we undersample and split,. Results will be better because of a fictitious dataset with data that seems more balanced. So split and do balance on the training dataset. The real world may not perform in the way we expect it to. But maybe there’s something we can learn.
  + Amit: did an assignment to predict something and it didn’t really work. Even so, have tools and methods to examine what’s going on.
  + Amanda: we’ve split between train and test and balance training and predicting with true. It seems that’s the correct way of doing but Amit was saying we balance and then split it. Because we’re sort of faking the data and we get better results, but is it a true model.
  + Iris: Was balancing and then splitting. It does seem like that’s not the real dataset and Amanda’s justification makes sense. So balance it in the training and evaluate on the test set.
* Sampling
  + Ivan did some oversampling. He got better performance with oversampling, but it might be overfitted.
  + Decided not to do oversampling
* Baseline model has been added before Kmeans
* Review powerpoint slides and distribution
* Analysis of models
* Any other questions
* Next steps:
  + Finalize Master notebook
  + Powerpoint distribution (15-20 min for each group)
  + Schedule run through - Monday, Dec. 2nd, 8pm

Hi all, Below are our next steps as we begin to finalize our project:

* Finalize the Master notebook (merge your models onto this notebook with Amanda’s pre-processed dataset if you haven’t already, or work with other folks to merge models) by Friday, Dec. 2nd
* Finalize powerpoint slides by Friday, Dec. 2nd
* Run through of presentation will take place on Monday, Dec. 5th at 8pm EST / 5pm PST

## Meeting on 11/27

Agenda:

* Update on Uri’s OH
  + Play around. Inbound dataset, what we should do.
  + 3 options:
    - 1. Undersampling - Do HW10 method, take equal number from every genre - Iris suggests this. 80% from least popular and having an equal number of that throughout
    - 2. Drop the lowest -
    - Iris tried it and no difference
    - 3. Use the same proportions for training and test dataset.
  + Dedupe - single or multiple classifications, Uri suggests dedupe so that there’s only one of each.
    - If we do only one classification, accuracy drops a bit.
  + Why imbalanced data? Results weren’t as good and feedback from Amit. Random forest went from 30 to 70%.
    - Amanda has been working on this and has been using under sampling (take a proportion of that, like HW10). Another approach is oversampling takes a smaller genre and uses synthetic data points and will create data points between those two records (SMOTE). It’s pretty easy to apply, but requires a pip install. We can show each model with different strategies.
    - Uri discourages SMOTE because it’s synthetic data. However, synthetic data can pick up some patterns and intended to not mess up balance of data and bulks up the scale a bit.
* Update from Shuo & Iris
* How should we pre-process our dataset (have we been consistent on what columns/values we decide to drop from our models)?
  + Look at how we balance
  + How we scale
  + We should draw from the same features and allow different iteration sof models. We can drop anything unnecessary for that model.
  + Minmax scaler? Used it because MinMaxScaler will transform each value in the column proportionally within the range [0,1]. This is quite acceptable in cases where we are not concerned about the standardization along the variance axes. (E.g. image processing or neural networks expecting values between 0 to 1).
* Discuss Analysis
* To do:
  + Consolidate notebook
    - Has everyone uploaded their notebooks (doesn’t need to have everyone’s we can settle on whose we think is best, maybe Shuo’s?)
    - Ivan to include his in your Master and upload it or needs some help notebook
  + Complete PPT slides
  + Plan presentation run through

Meeting updates:

* Review team meeting notes here.
* Preprocessing & balancing dataset:
  + To balance data, we have decided to go with the undersampling method (based off of Uri’s recommendation)
  + To keep our dataset consistent and uniform, Amanda has created a new Master notebook that standardizes how we pre-process our data (E.g. deduped, how we balance data, etc.). We will be able to use the same dataset for all our models and evaluate if there are improvements to accuracy scores when we test with different scaling/normalization methods (E.g. will Kmeans improve with minmaxscaler vs. standard scalar?)
  + Amanda is working on transferring our models from our individual notebooks into the new Master notebook. We will then review and edit our models we were responsible for and make any necessary changes.
* To do’s:
  + Amanda will do her best to transfer our notebooks.
  + Team will review and edit models accordingly by Wednesday
  + Ivan - can you upload your notebook, if you need assistance with the models, Amanda said she’d like to help!
  + Next meeting will be after class on Wednesday
  + Aim to complete our models/notebook and powerpoint by Dec. 2nd.

## 

## 

## Feedback from Baseline Submission 11/02

* What’s the baseline model? Suggest to predict the most common song in the dataset
* What’s going on with KNN (n=1 should always predict 100%)
  + KNN doesn’t have to be 100%
* What to do with an imbalance dataset (e.g., pop)?
* Iris: Once we address these the imbalance question, should we make a final clean dataset and rerun all our models based on the new cleaning?

## Meeting on 11/16

Agenda:

1. Consolidating notebooks (EDA & different models) into one
   1. How should we pre-process our dataset (have we been consistent on what columns/values we decide to drop from our models)?
2. Give an update on progress on ’What’s pending?’:
   1. Outlier analysis
   2. Complete the development of the K-Means model
      1. Have a look at it - is it right?
      2. What amount of clusters should we use?
         1. We set the variance threshold -
   3. Improve accuracy on some of the models
   4. Debug Neural Network model
   5. Compare and analyze models
      1. Tabular data and can’t be done. Ivan will keep working on alternate strategy (using same songs and adding more data).
3. Review FINAL Presentation and plan:
   1. Feedback on Baseline Submission
   2. Discuss if we want to aim to get better accuracy on models. If yes, how. -
      1. What Ivan is already testing
   3. What to do with songs with more than one genre classification - OH w/ Ishaani at 11am PST (Ivan) and Uri at 11:45am PST (Iris & Ivan)
      1. Randomly choose one genre?
      2. Run through same algorithm multiple times?
      3. Imbalance data - do we keep the same proportions in training and test or do HW10 (referred and not referred and get 80 from each, get 160 and split to test and validation. Taking x number from each of the labels) method?
   4. CNN - same as above
4. Next steps:
   1. Final presentation - assign parts
5. Discuss team’s availability in the next weeks
   1. Iris will be out the Friday before.
   2. Next meeting on Nov. 27
   3. Ivan upload notebook by this weekend
   4. Internal deadline to be done w/ ppt on Dec. 2nd (and can do run through on 3rd 4th)

## 

## Meeting on 11/01

Agenda:

* Run through baseline presentation
  + Time it and record it.
  + Amanda will be out of class tomorrow.
* Tasks:

## Meeting on 10/28

Agenda:

* Choose a dataset - go with the old dataset.
  + Outliers (Ivan) - standard form of working it. Following a recipe, but if we analyze the data, it varies by model, thats why so much data seems to be out of the average. That could be a reason why
* Discuss model options: test for accuracies and see which are the best!
  + KNN (Daisy)
  + Random Forest (Iris)
  + Gradient Boosting (Ivan)
  + Neural Network (Ivan)
  + Logistic regression (Iris)
* Tasks: slides done by Monday, 10/31. We’ll claim slides by Monday’s end?
  + Distribute slides
  + Baseline presentation (who will present what?)
* Confirm next meeting: Tuesday?
  + Iris unavailable from 6:30-8 PM PT

UPDATES:

* We will be going with our original (old) dataset.
* We will be comparing models and test accuracies to see which model is best. Below are the models we’ll be comparing with each other. If there are any models you’re interested in helping with, feel free to reach out to the person working on it!:
  + Kmeans - completed but need to tweek (Daisy)
  + Logistic regression - completed but need to tweek (Iris)
  + KNN (Daisy)
  + Random Forest (Iris)
  + Gradient Boosting (Ivan)
  + Neural Network (Ivan)
* Deadlines:
  + 1) Baseline presentation slides to be completed by Mon. Oct. 31.
    - Slides 2-3 (Intro & Objective) Shuo can you complete these slides? (these may already be completed, so you may want to help Amanda with some slides?)
    - Slides 4-7 (EDA & Findings) Amanda, since you did a lot of the EDA, can you complete these slides?
    - Slide 8 (Outliers) - Ivan will work on this slide
    - Slide 9 (Model) - (Iris, Daisy, Ivan, and the rest of the team) complete baseline model slides if completed by next week.
  + 2) Meet on Tues. Nov. 1 at 5:30pm EST/ 2:30pm PST to do a baseline presentation run through

## Meeting on 10/12

Agenda:

* Discussing what each of us did for EDA
  + Finalizing an EDA notebook
* Finalize problem statement/focus & model:
  + Ideas:
    - predicting song genres based on song characteristics (classification)
    - Song recommendations based on genres & characteristics
    - cluster analysis and song duration prediction (E.g. will a techno song be longer than a pop song, etc?)
    - song names and song genre prediction (Ivan’s idea mentioned in our last meeting)
    - Genre prediction based on song characteristics (multiclass classification, etc.)
      * We can also create clusters & kmeans
      * Song title - We can use it to multiclass classification and see if e.g. more rnb songs have ‘love’, etc. We’d also have to use smaller dataset. (Non english songs in dataset also)
      * Business case can be running through algorithm and classify it for spotify internally
    - Song recommendation (KNN, kmeans, etc.)
      * Would a customer might want a similar song played next? May not recommend the exact same but something close.
        + To respond to question ^^, we can frame our business scenario of the next closest song
      * If recommend a different genre and making recommendations outside of genres.
      * What would be our measure of success with song recommendation?
    - Song titles
* Discussing what we need for baseline submission
  + Powerpoint
  + Start Phase 1 model
* Task assignments:
  + Start Phase 1 model
    - Everyone works independently and tackles the model section - **Everyone by 10/19**
  + Prepare EDA notebook for baseline presentation
  + Powerpoint - **Daisy start framework by 10/19**

## Meeting on 10/05

Time zones:

* Daisy & Iris - PST
* Amanda, Ivan, Shuo - EST

Agenda:

* Select dataset we’re going to use
  + Refer to this spreadsheet for suggested ideas: [Datasets](https://docs.google.com/spreadsheets/d/1D_mIx8irgyyP4M7llxxG9CreHqaP62nbON2VmiVyS_A/edit#gid=0)
  + Going to use the Spotify dataset. Use genres\_v2.csv
* What’s our problem statement/business case?
* Task assignments:
  + **For the Baseline Submission:** 
    - EDA (summary statistics with basic histograms, count of missing values, visualization, etc.) - Everyone by 10/12
      * We’ll all periodically share/update our .ipynb files on the shared Google Drive
    - Choose ML model and method - by 10/12
  + **For the Final Presentation:** 
    - Data cleaning
    - Confirm model and ML method
    - Prepare data (shuffle data, split data into train and test)
    - Build model
    - Train model & evaluate (fit, predict, etc.)
    - Results & Analysis (model summary, model explainability, etc.)
* Target deadlines:
  + Teams’ availability?:
    - Daisy away Oct. 17-26
    - Iris unavailable on Oct 22, 23, 29.
    - Amanda Oct. 18-24
    - Shuo unavailable on Oct 25, Nov. 11-13

## Baseline Presentation Guidelines:

