**Senior Data Scientist Challenge**

## **PART 1**

**Music Classification Mini-Project:**

A client has requested us to build a composer classifier based on live captured audio that can be set to stream chunks of data to our model at 15, 30, 60-second intervals, as desired. For this project, the client has collected and annotated several hundred audio files and have saved them as simple midi files. The provided PS1 Folder contains the known midi files pertaining to four (4) composers: Bach, Beethoven, Schubert, and Brahms.

**Goal:**

The goal of this project is to develop a classifier/pipeline that is able to determine which midi files in the provided PS2 folder are not written by the four (4) composers above (it is a small number).

**Details:**

In the interest of time, 30-seconds is the desired default setting. Analysis of how data quantity effects performance is not expected, but will highlight important client requirements.

The midi files contain additional information that may not be available at inference time; ensure that your algorithm can support this.

The name of the file before the first underscore is the composition name.

The attached dataset is provided in midi format and taken from Musicnet.

(<https://arxiv.org/pdf/1611.09827.pdf>).

Optimal model performance is not expected, and neither is the use of the entire dataset and information. Classical ML approach(es) are recommended as the main form of analysis as setting up a thorough pipeline will be evaluated by the client more favorably than a poorly implemented state-of-the-art model that is not well-validated or documented. A deep learning approach may be attempted based on results to showcase additional capabilities.

The client exercise is geared towards a truncated, real-world scenario, demonstrating: (1) Efficacy: How fast you can build solutions and the quantity of analysis shown; (2) Quality: How well you showcase your understanding of the problem, data, and code; (3) Robustness: The completeness of your analysis flow and the robustness of the validation process.

In particular, the client will consider overall problem understanding and set up of the analysis; the completeness of your EDA and model building/tuning/validation; result interpretation and analysis approach and outcomes; code quality and completeness; and final conclusions and recommendations based on the brief sprint.

## **Using Open-source Libraries/Packages:**

Any open-source code/libraries and publicly available information is encouraged providing you cite the material.

**Return & Due Date:**

Please email back the final notebook within 72 hours of receiving this exercise. Please return a brief abstract of the solution and findings. **For ease of review, please save your analysis in both an HTML file as well as ipynb.**

“

**PART 2**

# **Pandas Basics:**

The dataset for this part of the interview can be found at [Link](https://www.dropbox.com/s/niqs8pv5n55kwu0/Coding%20DS%20dataset.zip?dl=0), You will see two .csv files which will form the basis of this part.

Please use a Jupyter notebook and perform the following manipulations in order.

1. Concat: vertically concatenate the two csvs into a new dataframe **df\_full**; for missing values in the test set use a random selection of values in the same column from the train set.
2. Filter: on **df\_full** select only the weekdays, where the relative humidity is >50 and abso <1.4 denote this **df\_filtered**.
3. Apply: on **df\_filtered** perform sensor 2 \* sensor 3 if sensor 4 < MEDIAN VALUE, else sensor 2.\* sensor 4 - average of that calendar month value
4. Groupby: on **df\_fill** table find the min/max/mean/count temp over each day as a table
5. Create a column on **df\_full**:
   1. read in a dict : 5 unequal splits, 0-0.2, 0.2-0.5, 0.5-0.8, 0.8-0.85, 1 - make it generic
   2. apply to sensor 1 = name it Truth, apply to sensor 2 = name it Predicted
6. Merge: starting from **df\_full** let’s look at the differences to the same day last year for available days
   1. How many days overlap in the two latest years?
   2. What is the average of the daily difference of abs humidity?
   3. What is the difference in average of sensor 1 data between the two years?
7. Plots on **df\_full**:
   1. histogram of absolute humidity
   2. sensor 1 - 5 histograms overlaid on same figure
   3. scatter plot of sensor 1 vs sensor 2 coloured blue if weekday, red otherwise
   4. correlation matrix of all sensors against each other with histograms along diag
   5. confusion matrix of predicted vs truth
8. Create data split: 70%/30% split of the data from the **train.csv** ensure that each day is not split between sets

# 

**Return & Due Date:**

Please email back your submission within 72 hours of receiving this exercise.

**“**As discussed, the technical team will evaluate your code and results, and invite you to present on your background, methods, and results. I will provide additional information on the presentation when scheduling. In any case, I will email you after we review your submission and provide any additional feedback and next steps.

Thank you for taking the time to complete this exercise; we appreciate your time, effort, and attention upfront. Eager to see your work and good luck in your analysis.”

**REVIEW**

Notebook and submission will be returned to Camille Cramer, Alexander Tolpygo, and other designated reviewers in the original format.

Assessment response will be provided with 24-48 hours depending on timing and current activity.

Since the role as Senior Data Scientist requires the combined skills of technical execution, critical thinking, and clear communication, the following will be considered:

* Technical knowledge (show relevant and deep knowledge on the topic through problem solving, EDA, modeling, analysis volume, code quality)
* Analytical Skills (showcases logic and analysis robustness)
* Communication (shows clarity in approach, formatting, communication, response to conclusions or recommendations )

**POST-CHALLENGE; NEXT INTERVIEW REQUEST**

**[To be sent to Senior Data Scientist Candidates should they pass the technical challenge]**

“  
I hope you are well since the last time we spoke and thank you again for the time you have invested into the interview process. The team has reviewed your work and would like to move into the next step of the interview process—congratulations!

As previously mentioned, we would like to invite you to present your work where you will break down your analysis through a technical discussion as if presenting to a client technical group. In terms of presentation structure, please prepare the following:

* 5-10 minutes introduction, please include any relevant previous work and academic experiences
* 20-25 minutes discussing the project exercise, including presenting the approach, methods, conclusions, and recommendations
* Open Q&A and technical discussion

During this one hour, the team would like to discuss general AI-related questions about your experience, capabilities and approach to the project.

Please be available for at least 60 minutes for this stage.

**TECHNICAL PRESENTATION & QUESTIONS:**

**Deloitte reviewers will ask general and technical questions and prompts based on the presented work and model selections.**

**Deloitte reviewers will evaluate presentation delivery, content, thoughtfulness of response, technical understanding, and other background information in reference to listed resume experience.**

**This may include, as demonstrated by the candidate knowledge, the following questions related to the analysis:**

**Example Basic Machine Learning Questions**

1. What is overfitting? How do we avoid it?
2. What is bias/variance and how do they differ?
3. What is data splitting, what is its purpose?
4. How would you split your data and what are common problems you might have? Are there additional training techniques and auxiliary methods you would use to mitigate any data issues if:

* 10 subjects with 50-100 time series of lengths between 30 seconds and 10 minutes. Outcomes are balanced between subjects. Classification.
* Twitter data: Tweets and replies on 1 million original tweets. Sentiment analysis.
* EHR’s from 20k patients spread across 12 hospitals and 200 doctors with skewed distributions (80% of patients come from 1 hospital). Fraud detection.

1. What is K-fold cross validation? Leave one out cross validation? How do they work and when would you use them?
2. How would you split your data for unbalanced classes?
3. What is supervised vs. unsupervised classification? Examples of each and when are they relevant?
4. Deep learning is commonly used in our solutions. Please go through and explain the common components in an MLP deep learning model, how it is trained and what is the purpose of each component. Explain the following components of a neural network and their purpose:
   * MLP/FC layer
   * Loss function
   * Back propagation / Optimization Rules / Name a few
   * Activation function / Name a few
   * Regularization
   * Hyper Parameters / Name a few
   * Mini Batching
   * Training / Inference Mode differences