Data Mapping

> *... revisiting solutions that can read data in CSV files (with and without headings) from clients and map each record to standard headings in pControl*

Next Steps

- [x] stand up the previous solution (Alpha from Goodbits vendor)
- [x] evaluate the solution
- [x] confirm the solution results
- [ ] confirm the next development steps to implement the expected functionality
    1. describe the current workflow for client onboarding and identify where is the "big payoff"
    2. codify the workflow for client onboarding
        a. redesign the solution for production use, e.g. redesigning the Alpha solution with the correct use of "out-of-the-box" solutions (Ludwig, Amazon Comprehend, Amazon SageMaker, and other available ML toolboxes)
        b. release the redesigned solution into production
        c. confirm the improvements to solution in production, e.g. adding functionalities of Data Integrity for market data (formulas)

Alpha Solution

- this solution was expected to understand the data in client onboarding files, match each data point to pControl data category names, and improve on results by user interaction
- this was a data mapping exercise to design a solution that learns patterns in data and increases the accuracy of match with more incoming data

result

- a solution assigns the standard heading name from a predefined business dictionary to each group of records (columns) in a file
    - it also highlights standard heading names that are missing in provided files (for files without headings)
- a business dictionary was created by MG to hold few possible heading names (frequently occurring in client files) for each standard heading name in pControl
- results are available on the browser screen (frontend app)
    - results can be downloaded as a file for files without headers only
- this solution is NOT learning from user interaction (users are feeding new data) and its performance will NOT improve since it does not have any learning component implemented
    - for files with headers, the type of implemented algorithm is not possible to train and its performance depends on a quality of predefined business dictionary
    - for files without headers, the implemented algorithm is trained only once with an unknown training dataset and then it predicts the results from the same trained model with a poor performance
- training data for the text classification algorithm (files without headers) were NOT disclosed
    - labeled data in test_1.csv were not provided
- instructions to completely standup the solution were not provided
    - backend app instructions were provided
    - frontend app instructions and instructions for a connection between backend and frontend were not provided

live demo

- feel free to access the Alpha solution via this URL when you are not on VPN
- this solution is live on MG AWS cloud since 08 Sep 2021
- ⚠️ **make sure to use public data only, NOT production data**

code repository

- feel free to access the Alpha repository via this URL when on VPN
- this repo was created on MG Bitbucket Server on 02 Sep 2021

technology

machine learning	backend app	frontend app

text similarity task on files with headings • implemented via Python library (FuzzyWuzzy https://github.com/seatgeek/fuzzywuzzy) • Levensthein distance method to match strings • it shows the highest score of match (100 % if perfect match) • this solution is based only on the headings in file (not on records / values for each heading)	• developed in the Flask web application framework • coded in Python • monitored by Sentry	• developed in the Angular web application framework • coded in TypeScript and JavaScript
text classification task on files without headings • implemented via ML toolbox (Ludwig https://github.com/ludwig-ai/ludwig) • parallel CNN (Convolutional Neural Network) neural net • it shows the probability of top 3 columns in files for each standard heading • it highlights missing standard headings if the sum of top 3 column probabilities is less then 10 % • this solution is based on the first 10,000 records / values in file		

data

business dictionary	training set	test set
business_...onary.pdf	• not available • obfuscated client data that are labeled and split into a training and validation sets	• obfuscated client data • u100 files with and without headers