Fun with Data

April 22, 2021

Overview

1. Data Exploration
2. Working with a client
    a. Application Design
    b. Data modeling and prediction

Part 1 - Data Exploration

Exploring the “Federal Employee Viewpoint Survey” by the US Office of Personnel Management

- Basic statistics
- Data quality assessment

Part 2 - Task A - Application Design

Design and implement an application for the US Office of Personnel Management

- Questions
- Assumptions
- Diagram of one option

Part 2- Task B - Data Modeling and Prediction

Model design for the survey data

- Feature treatment
- Model choice
- Model performance

Part 2 - Task A - Application Design

Design and implement an application for the US Office of Personnel Management

- Questions
- Assumptions
- Diagram of one option

Part 1 - Data Exploration

Exploring the “Federal Employee Viewpoint Survey” by the US Office of Personnel Management

- Basic statistics
- Data quality assessment

Part 2- Task B - Data Modeling and Prediction

Model design for the survey data

- Feature treatment
- Model choice
- Model performance

Part 2 - Task A - Application Design

Scenario

How would you design and implement an application for a client when the client would like to predict what departments will experience significant turnover.

Questions to Answer

There would likely be a discovery session with the client.

1. How do they hope to action the prediction of the significant turnover in a department?
2. What restrictions should we be aware of?
3. What data is available for the application?
4. etc.

Assumptions

Lacking a discovery session the following assumptions are made.

- Survey data is not real time
- Data must be stored on the order of several years to provide trend reporting and model improvement
- Data regarding is available
- Survey is provided by third party data collection
- Existing infrastructure can be leveraged
    a. AWS, Data integration vendor, etc.
- Survey data does not currently need to be available with data from other systems
- Cost is not a problem

Assumptions Continued

Assumptions

- Survey data is not real time
- Data stored over several years

Implications

- Streaming solution not needed (sorry Kafka)
- End state of the data can a RDBMS (MySQL, PostgreSQL, etc.) but S3 is lower maintenance

Assumptions Continued

Assumption

- Data regarding actual turnover is available
- Survey data does not currently need to be available with data from other systems

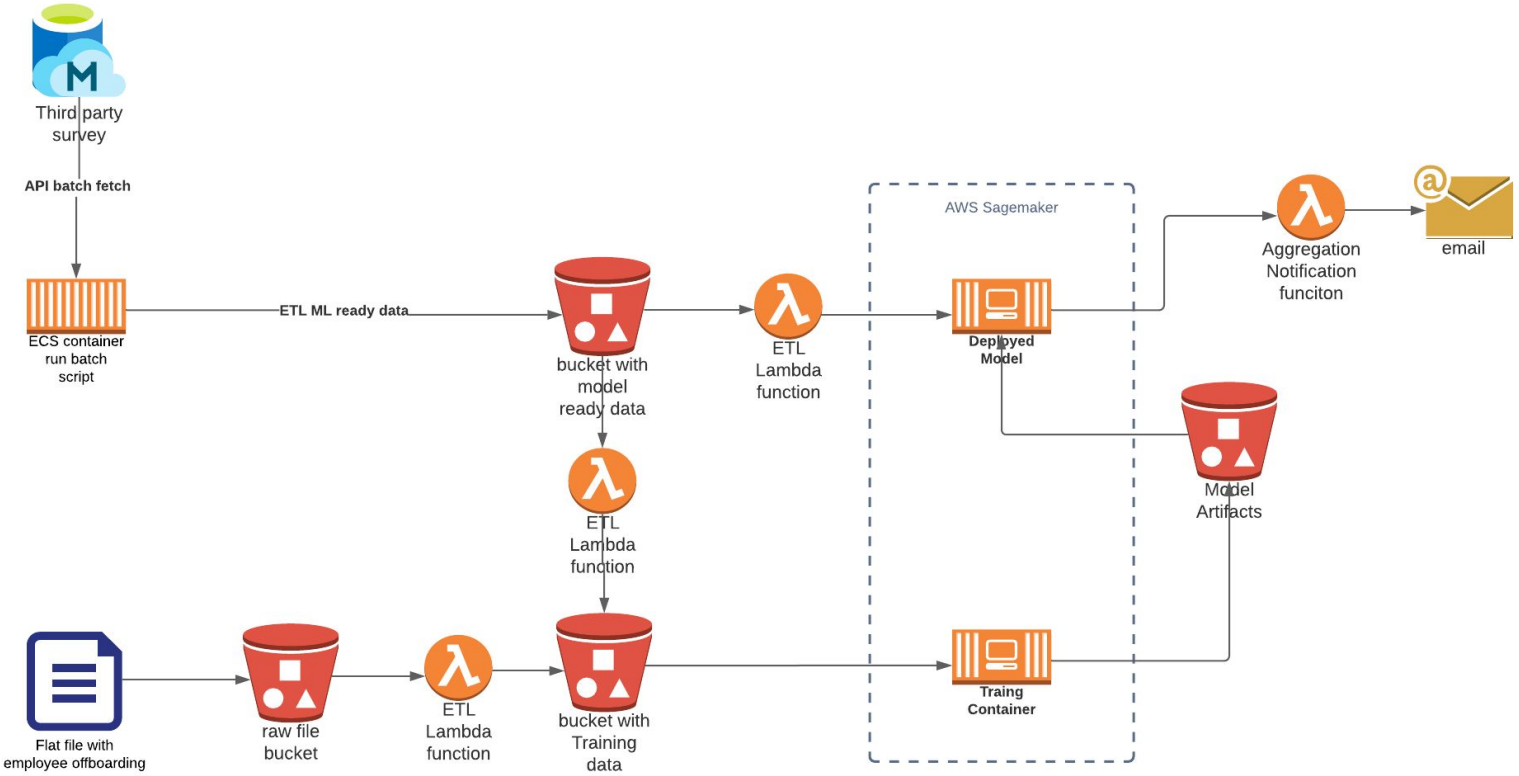
Implications

- Som combining of data is needed
- Although, currently no need to incorporate other data, the data has most flexibility in S3

Item 1

Item 2

Application Diagram



What's Missing?

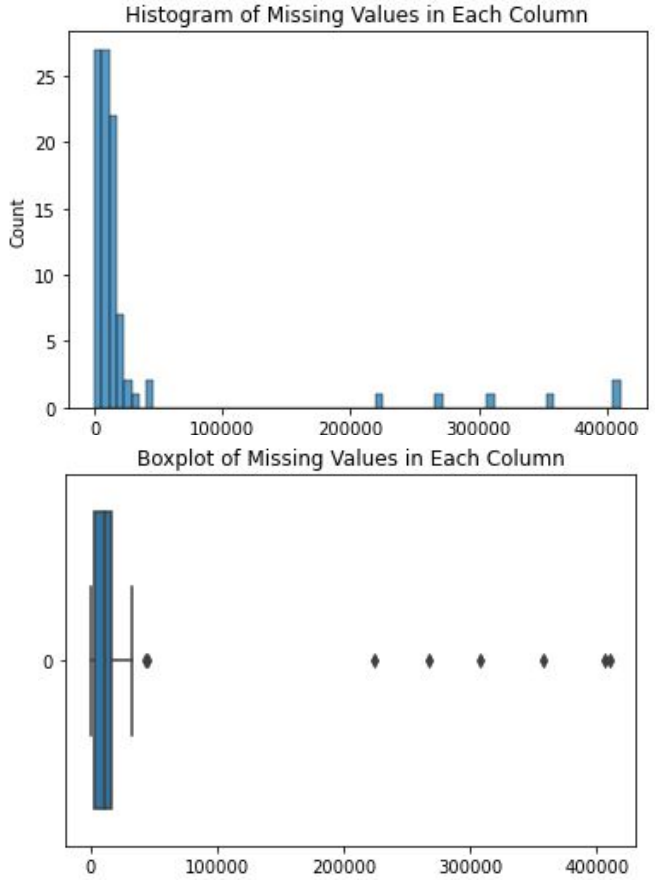
- Logging
- Deployment method (infrastructure as code)
- Long term vision
- Methods for analysts and scientists to access the data
- Existing data integration vendor considerations
    a. Not reinventing the wheel
- Actual Termination data
- Data management strategy (GDPR)
- Cost concerns

Part 1 -
Data Exploration

Basic statistics

First impression - Lots of missing data!

1. Missing ~7% of all potential data
2. Average missing values per column: 31270
3. Standard deviation:80550
4. Many outlier columns



Data Quality Assessment

1. A few columns with a majority of no data
2. A few Mixed type columns
    a. Float columns with Encoded categories
3. Object columns are str except where null
4. Random ID are unique with no null

Part 2- Task B -
Data Modeling and Prediction

Feature Treatment

This is survey data, majority categorical data

1. Demographic columns were str
    a. Others pre Encoded
2. One Hot Swap used on object columns
3. Random ID set as the Index
4. ‘DLEAVING’ column set as our target column to predict
5. 80/20 used on 80% of data for Training(64%) Validation(16%) and Testing(20%)

Model Choice

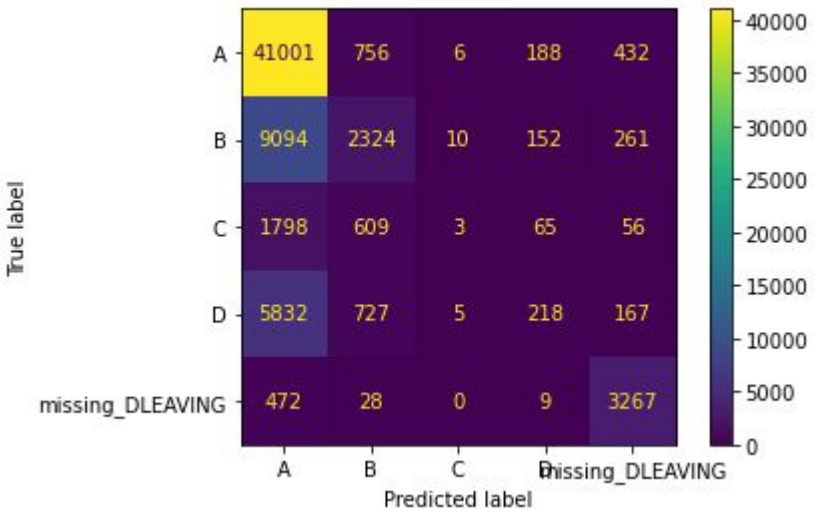
SGDClassifier

1. Large amount of data >100k
2. Overwhelming choice of one answer
3. Processing time

Model Performance

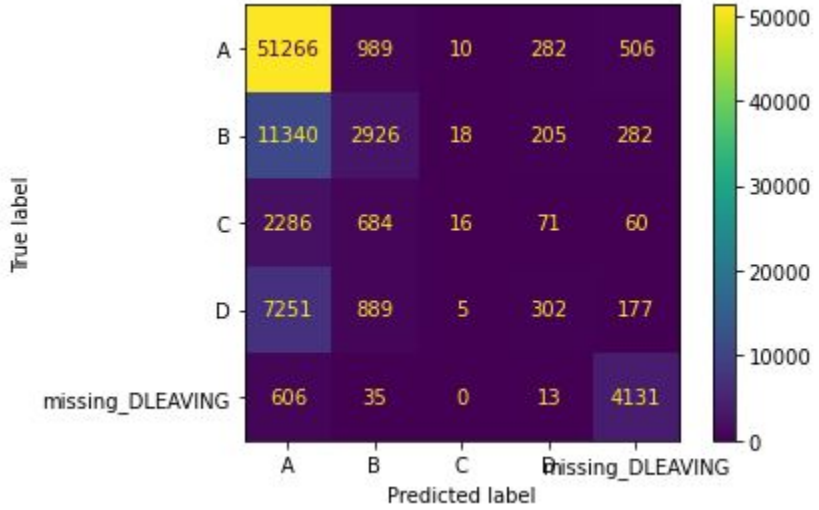
Validation

69.37% Accuracy



Test

69.52% Accuracy



Questions