GLENTEL

# “Identifying and Analyzing Traits Associated with High Performers

Clara Su, Manuel Maldonado, Robert Pimentel and Thomas Pin - **Mentor**: Varada Kolhatkar

05/08/2020

# Glentel

- **Mobile Phone Retailer in Canada**
- Joint venture between Rogers Communications and Bell Canada Enterprises
- 350 locations spread over three banners
- 2,000 employees
    - Sales Associate
    - Assistant Manager
    - Sales Manager


WIRELESSWAVE
TM


booth
wireless™


wireless etc™


COSTCO®
WHOLESALE

# Data Science Problem

**HR wishes to understand what are the traits that are associated with employees who become high performers?"**

Why?

- **Decrease turnover rate** in new hires.
- **Optimize work force** in function of performance expectation.

## Current Solution

- No quantitative support for high-performing traits looked for in new hires.

# Data

Two sources of data:

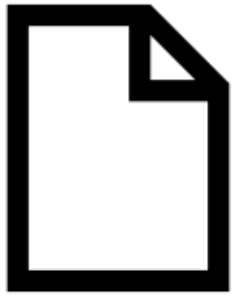
- **Unstructured data**
  - 400-600 Resumes (2019 onwards)
- **Structured (tabular)**
  - Employee demographics
  - Sales: phone/line “activation” data (2018 onwards)
  - Compensation tier based on activations
  - Termination reason
  - Promotions
  - Transfers

# Data Challenges

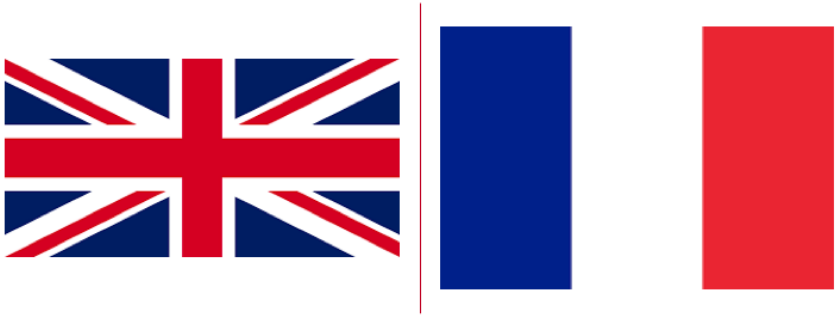
Style_1.pdf Style_2.docx

- Different file formats (PDF's, doc, docx, rtf, txt)
- Different text formats

# Data Challenges

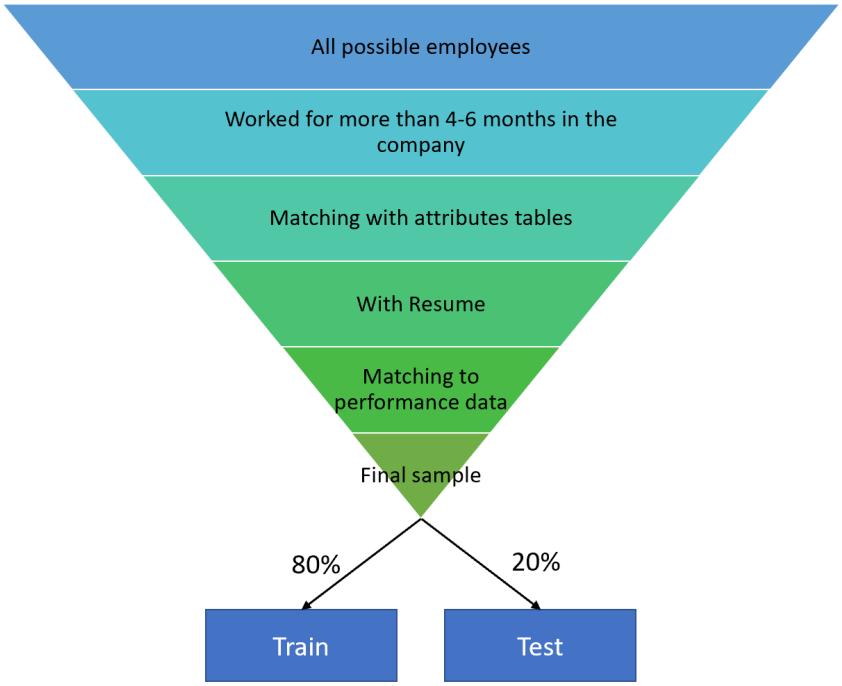


- Estimate 5%-8%* are blank (based off of a sample size of 150)



- Resume in both French and English

# Data Challenges



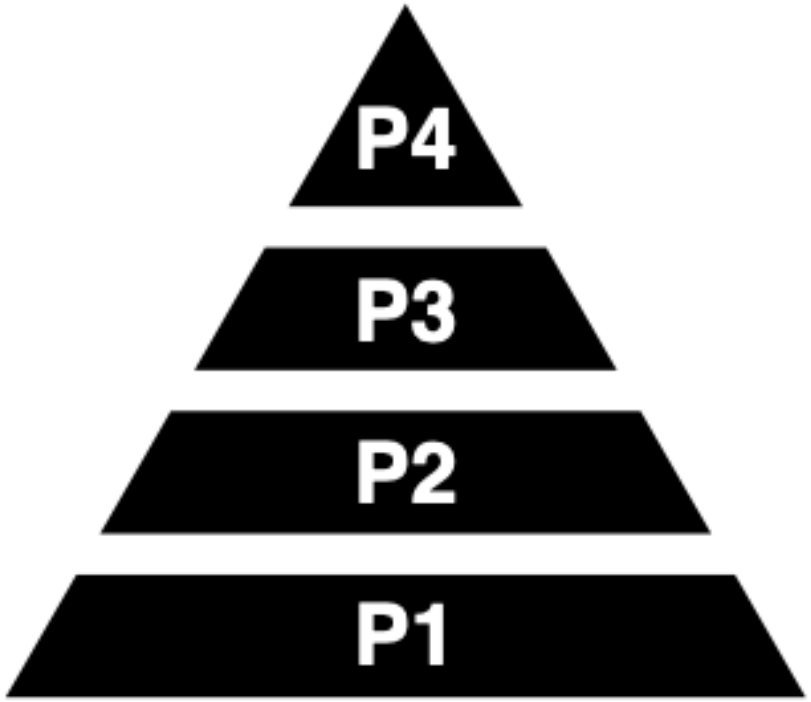

- Low sample size, limited primarly by resume data.

# Target (Response Variable)

- Performance Level: Binary
  - High Performer (1)
  - Low Performer (0)

- Requested by the company, based on their payroll tier reached.
- Influenced by sample size.

# Target (Response Variable)




- Pay achievement level

# Data Preparation: Ideal Data Format

| Employee_ID | Text | Feature2 | Feature3 | Feature_n | Target |
|---|---|---|---|---|---|
| **231** | … | 0 | 1 | 0 | High Performer |
| **456** | … | 1 | 1 | 1 | Non-High Performer |
| **790** | … | 0 | 0 | 1 | Non-High Performer |

# Potential Features

- **Hire Type**
  - Re-hire
  - Referral

- **Resume Features**

(Obtained through information retrieval techniques)

- Education
- Experience
- Job hopping
- Job experience (type and time)
- Spelling mistakes
- Language

# Our approach; using NLP

- Extracting text from resumes
- Pre-processing (stopwords, special characters, punctuation)
- Topic modelling.
- Feature engineering:
    - information retrieval
    - count vectorizer

# Feature Engineering

| Resume | College | Retail | Teamwork | Sales | Communication | Target |
|---|---|---|---|---|---|---|
| **Resume 1** | 1 | 0 | 1 | 0 | 1 | High Performer (1) |
| **Resume 2** | 0 | 1 | 1 | 0 | 1 | Non-High Performer (0) |

- We expect to mix count vectorizer features and engineered features through information retrieval techniques.
- Based on:
  - insights of topic modeling
  - partner's expertise.
  - data scientists' criteria/creativity.

# Machine Learning Pipeline

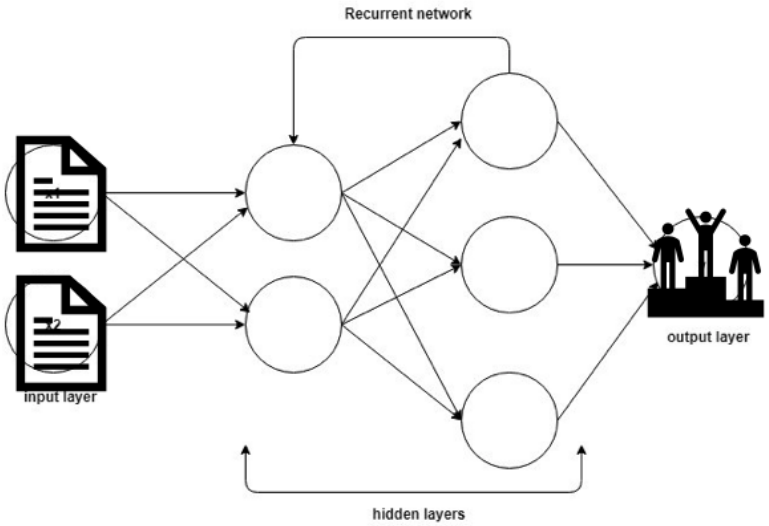
1. Feature Scaling
2. Cross validation
3. Hyperparameter optimization
4. Feature Selection
5. Model training
6. Model Interpretation

**Models**

- Baseline : logistic regression
- Ensemble of additional classifiers, such as:
  - SVM
  - Random Forest
  - Multilayer perceptron

# Complex Model Limitations

## LSTMs

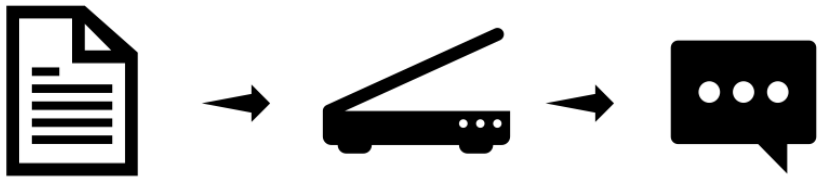



- Low sample size (overall aprox 400 observations)
- Deep learning models are hard to interpret.
- Data senstivity, can’t be uploaded to be cloud (names on resume text).

# Plan

**Week 1** May 4 - May 8

- Understanding/defining problem to be solved with Glentel
- Methodology research and definition
- Preliminary EDA on tabular datasets



- Partner still extracting resume information.

# Plan

**Week 2** May 11 - May 15

- Final Proposal to partners
- EDA

- Resume - Loading and Processing :
  - Special characters
  - Stop words
  - Lemmatization

# Plan

**Week 3-4** May 18 - May 29

**NLP**

- Count Vectorizer
- Topic Modelling
- Feature Engineering
    - Based on insights of the previous
    - Based on partner's expertise
    - Based on data scientist's criteria.

# Plan

**Week 5** Jun 1 - Jun 5

**Baseline Model Creation: Logistic Regression**

- Machine Learning Pipeline
  - scaling
  - cross validation
  - hyperparameter
  - Feature Selection
  - Measuring model Performance

# Plan

**Week 6-7** Jun 8 - Jun 12

**Challenging the baseline model with additional classifiers**

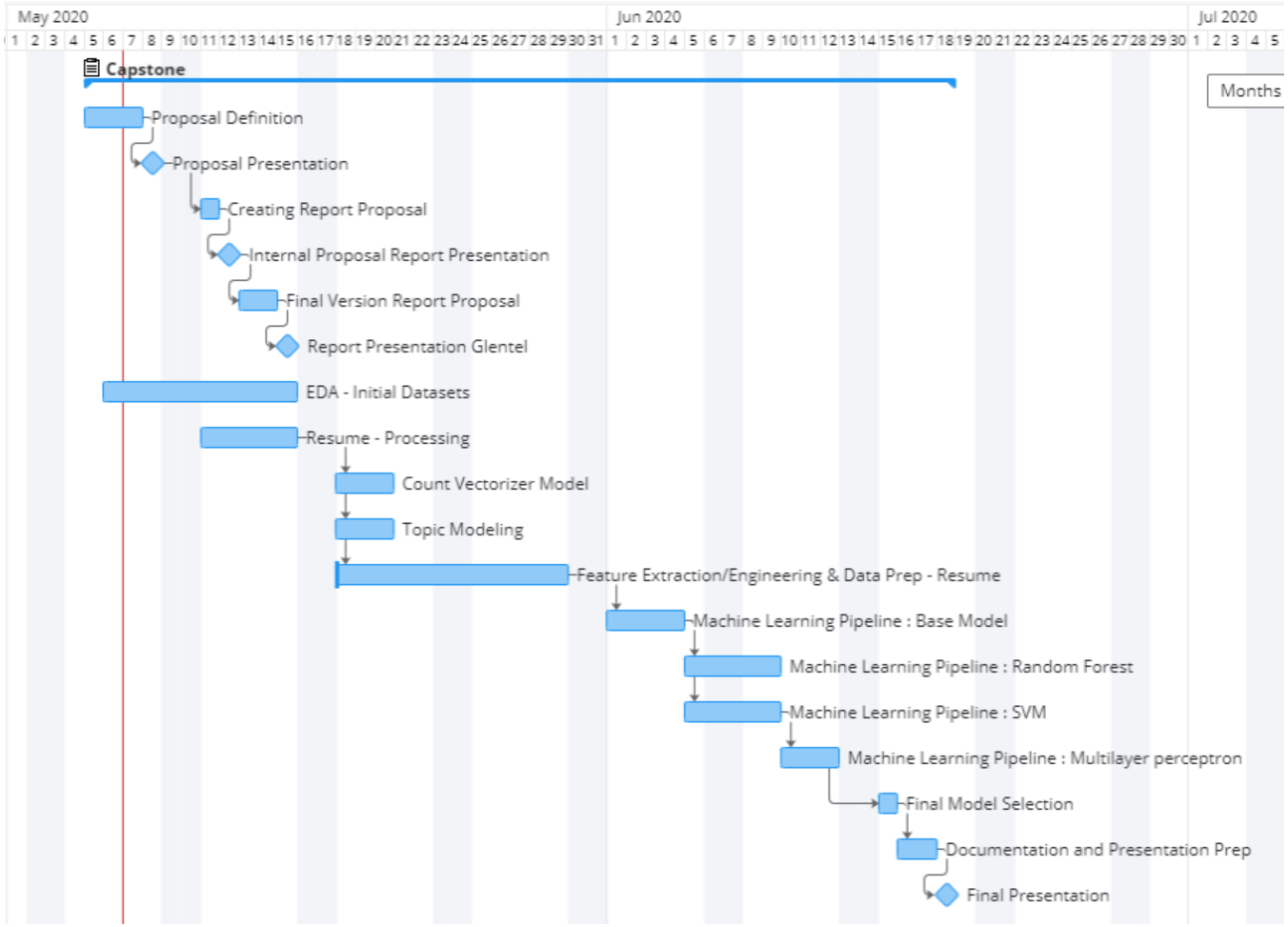
- Ensemble of additional classifiers, such as:
  - SVM
  - Random Forest
  - Multilayer perceptron

# Plan

**Week 8** Jun 15 - Jun 19

- Result documentation and comparison
- Final Model Selection
- Presentation preparation

# Gantt Chart


May 2020
Jun 2020
Jul 2020
Months
Capstone
Proposal Definition
Proposal Presentation
Creating Report Proposal
Internal Proposal Report Presentation
Final Version Report Proposal
Report Presentation Glentel
EDA - Initial Datasets
Resume - Processing
Count Vectorizer Model
Topic Modeling
Feature Extraction/Engineering & Data Prep - Resume
Machine Learning Pipeline : Base Model
Machine Learning Pipeline : Random Forest
Machine Learning Pipeline : SVM
Machine Learning Pipeline : Multilayer perceptron
Final Model Selection
Documentation and Presentation Prep
Final Presentation

# Questions?