Master of Data Science
Capstone Project

Glentel:
Identifying Traits Associated with High Performance

Clara Su
Manuel Maldonado
Robert Pimentel
Thomas Pin

Mentor: Varada Kolhatkar

Glentel

- Mobile phone retailer in Canada
- Joint venture between Rogers Communications and Bell
- Currently 2,000 employees:
  - Sales Associates
  - Assistant Manager
  - Sales Manager

Data Science Problem

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

Why?

- **Increase productivity in** sales performance
- **Decrease turnover rate** in new hires

Scientific Objective

Identify the traits correlated to high performance through the development of a predictive model.

Who is a High Performer?

An employee who reaches the business expectations in 75% of the time

- Business expectation is measured my monthly pay level reached

Example Employee

		Performance Window										
		Apr-19	May-19	Jun-19	Jul-19	Aug-19	Sep-19	Oct-19	Nov-19	Dec-19	Jan-20	Feb-20
Hire Month	Apr-19	0	0	1	1	0	1	1	1	0	1	0

1 = reached minimum expected compensation plateau during that month
0 = did not reach minimum expected compensation plateau during that month

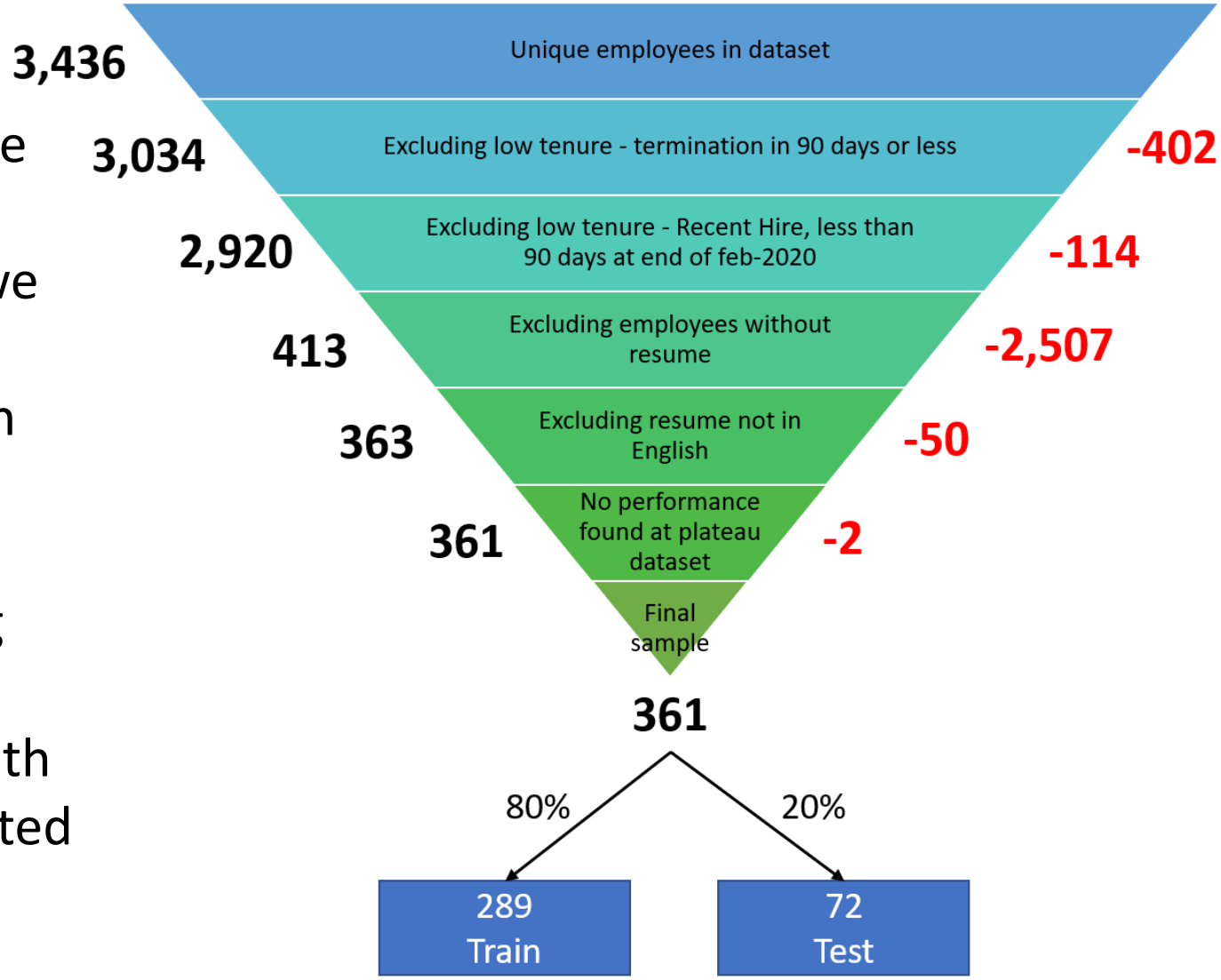
Months worked	11
Months excluding boarding process (first 2 months not included in count)	9
Months min plateau reached	6
High performance %	67%
Performance Classificaction	Non-High Performer

For the purposes of this analysis our definition was simplified high performance definition

Data & Resume Filtering

Considerations

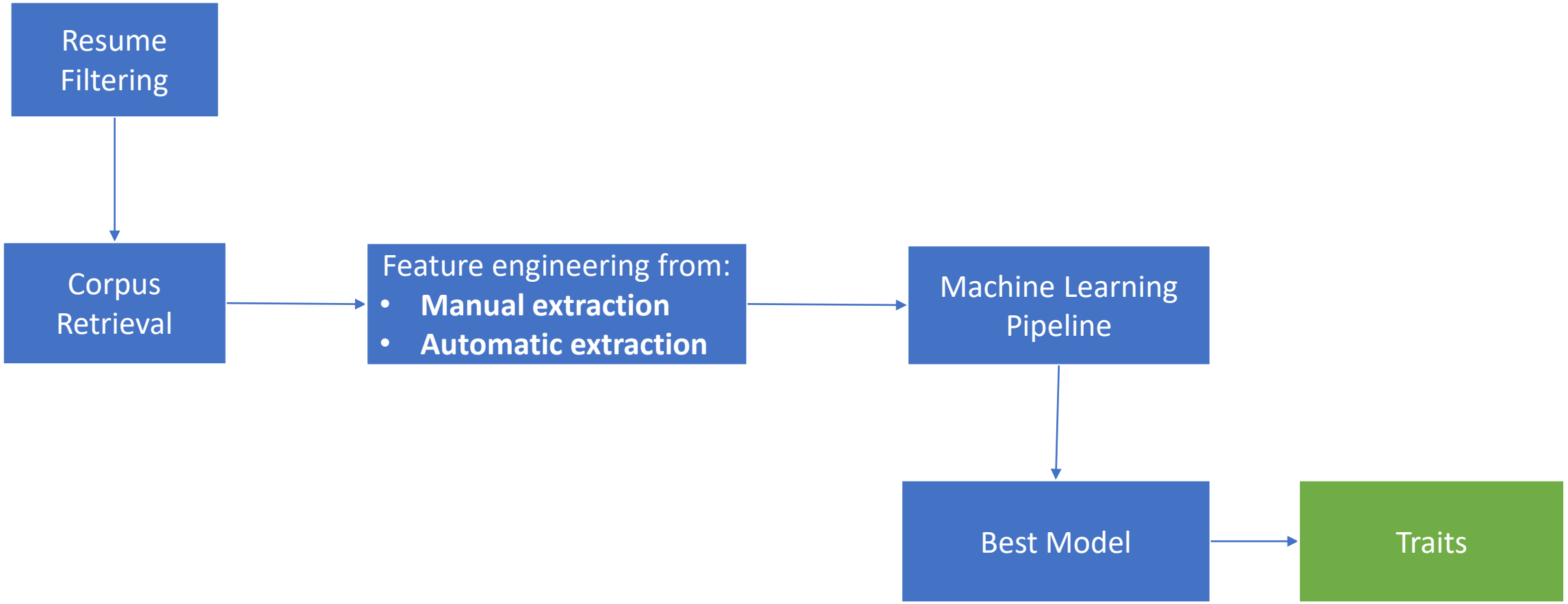
- Main source of applicant traits is the resume
- In order to measure performance we needed employees that:
  - Had worked at least 3 months in the company
    (The first two months are considered part of the boarding process)
  - Feb 2020 was the last month with “normal” performance, unaffected by COVID-19 Pandemic



Internal Definitions

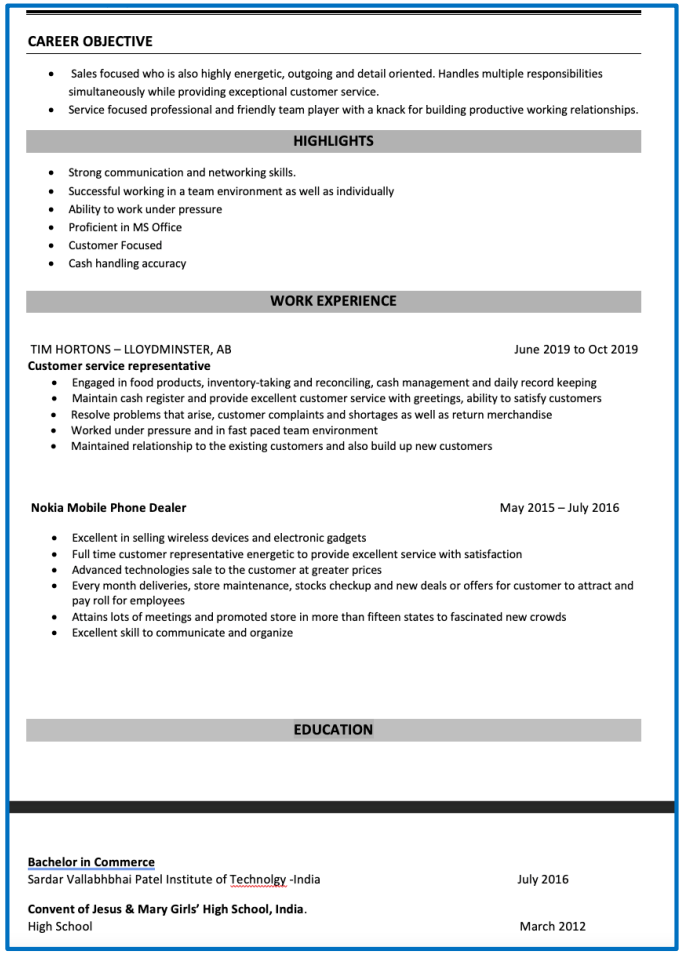
- Manual extraction: features engineered from manual extracted resume template
- Automatic extraction: features engineered from the resume corpus directly

Methodology



Corpus Extraction

Fictitious Resume Example



CAREER OBJECTIVE

- Sales focused who is also highly energetic, outgoing and detail oriented. Handles multiple responsibilities simultaneously while providing exceptional customer service.
- Service focused professional and friendly team player with a knack for building productive working relationships.

HIGHLIGHTS

- Strong communication and networking skills.
- Successful working in a team environment as well as individually
- Ability to work under pressure
- Proficient in MS Office
- Customer Focused
- Cash handling accuracy

WORK EXPERIENCE

TIM HORTONS – LLOYDMINSTER, AB — June 2019 to Oct 2019
Customer service representative

- Engaged in food products, inventory-taking and reconciling, cash management and daily record keeping
- Maintain cash register and provide excellent customer service with greetings, ability to satisfy customers
- Resolve problems that arise, customer complaints and shortages as well as return merchandise
- Worked under pressure and in fast paced team environment
- Maintained relationship to the existing customers and also build up new customers

Nokia Mobile Phone Dealer — May 2015 – July 2016

- Excellent in selling wireless devices and electronic gadgets
- Full time customer representative energetic to provide excellent service with satisfaction
- Advanced technologies sale to the customer at greater prices
- Every month deliveries, store maintenance, stocks checkup and new deals or offers for customer to attract and pay roll for employees
- Attains lots of meetings and promoted store in more than fifteen states to fascinated new crowds
- Excellent skill to communicate and organize

EDUCATION

Bachelor in Commerce
Sardar Vallabhbhai Patel Institute of Technolgy -India — July 2016

Convent of Jesus & Mary Girls' High School, India.
High School — March 2012

- Different file formats (.pdf, .doc, .docx, .rtf, .txt)
- Initial attempt to read data (1 wk)
  - **pdf plumber** (Slack Channel Recommendation)
  - **python-docx** (Win convert "doc -> docx" first)
- Final Selection
  - **Tika** (Can work with many formats)

Plain Text Extraction

resume_text
[redacted]CAREER OBJECTIVE · Sales focused who is also highly energetic, outgoing and detail oriented. Handles multiple responsibilities simultaneously while providing exceptional customer service. · Service focused professional and friendly team player with a knack for building productive working relationships. HIGHLIGHTS · Strong communication and networking skills. · Successful working in a team environment as well as individually · Ability to work under pressure · Proficient in MS Office · Customer Focused · Cash handling accuracy WORK EXPERIENCE TIM HORTONS – LLOYDMINSTER, AB June 2019 to Oct 2019 Customer service representative · Engaged in food products, inventory-taking and reconciling, cash management and daily record keeping · Maintain cash register and provide excellent customer service with greetings, ability to satisfy customers · Resolve problems that arise, customer complaints and shortages as well as return merchandise · Worked under pressure and in fast paced team environment · Maintained relationship to the existing customers and also build up new customers Nokia Mobile Phone Dealer May 2015 – July 2016 · Excellent in selling wireless devices and electronic gadgets · Full time customer representative energetic to provide excellent service with satisfaction · Advanced technologies sale to the customer at greater prices · Every month deliveries, store maintenance, stocks checkup and new deals or offers for customer to attract and pay roll for employees · Attains lots of meetings and promoted store in more than fifteen states to fascinated new crowds · Excellent skill to communicate and organize EDUCATION Bachelor in Commerce Sardar Vallabhbhai Patel Institute of Technolgy - India July 2016 Convent of Jesus & Mary Girls' High School, India. High School March 2012 REFERENCE Reference will be published upon request

Feature Engineering

Feature Engineering Techniques

Regex

- Exact match
- Multiple patterns
- Data cleaning
- Grouping & lists

Automating Communication Skills Proxy using Regex

Communication Skill words

```
1 wanna_capture_list  = ['communication skill', 'comunication skills',
2                        'communicationskills', 'comunicationskill',
3                        'comunicatorskil','communicator skills',
4                        'communicative skils']
5
6 regex_pattern = r"\b(comm?unic?[^\s]+ ?skill?s?)\b"
7
8 regex_test(regex_pattern, wanna_capture_list)
```

```
100.0 percent matched
NO MATCHED LIST: []
```

Segmented searching also created problems

Feature Engineering Techniques

Normalized Levenshtein Distance (Similarity Algorithm)

For example, the Levenshtein **distance between "kitten" and "sitting" is 3**,

1. **k**itten => **s**itten (substitute "k" for "s")
2. sitt**e**n => sitt**i**n (substitute "e" for "i")
3. sitt**i**n => sittin**g** (insert "g" at the end)

- Identify short strings in longer text with smalls differences
- No need to encode words into vectors such as Word2Vec
- Only 10 – 15% effectiveness

Feature Engineering (Industry Characterization)

Wikipedia company list
Levenshtein Only

Industry	Count employees	%
Unknown	920	85.3
Consumer Electronic	252	3.3
Clothing & Footwear	109	2.4

1079 companies

REGEX Key Words + Levenshtein

Industry	Count employees	%
Unknown	349	32.3
Telecommunications	252	23.3
Food Service	109	10.1
Sport_Travel_ Enterntain_Hotel	65	6.02
Food_Convenience_ Pharmace	65	5.46
Clothing & Footwear	59	2.68

Feature Engineering (Bag-of-Words)

	Neg feats	Neg weights	Pos feats	Pos weights
0	sale associate	-0.291163	communication skill	0.279128
1	high school	-0.191893	ensure customer	0.212689
2	customer service	-0.169689	customer satisfaction	0.198559
3	available request	-0.157930	sale target	0.196868
4	excellent customer	-0.137523	time frame	0.184704
5	health safety	-0.129824	high level	0.160940
6	team player	-0.127623	rogers communications	0.155917
7	store manager	-0.120927	hindi punjabi	0.155715

BOW

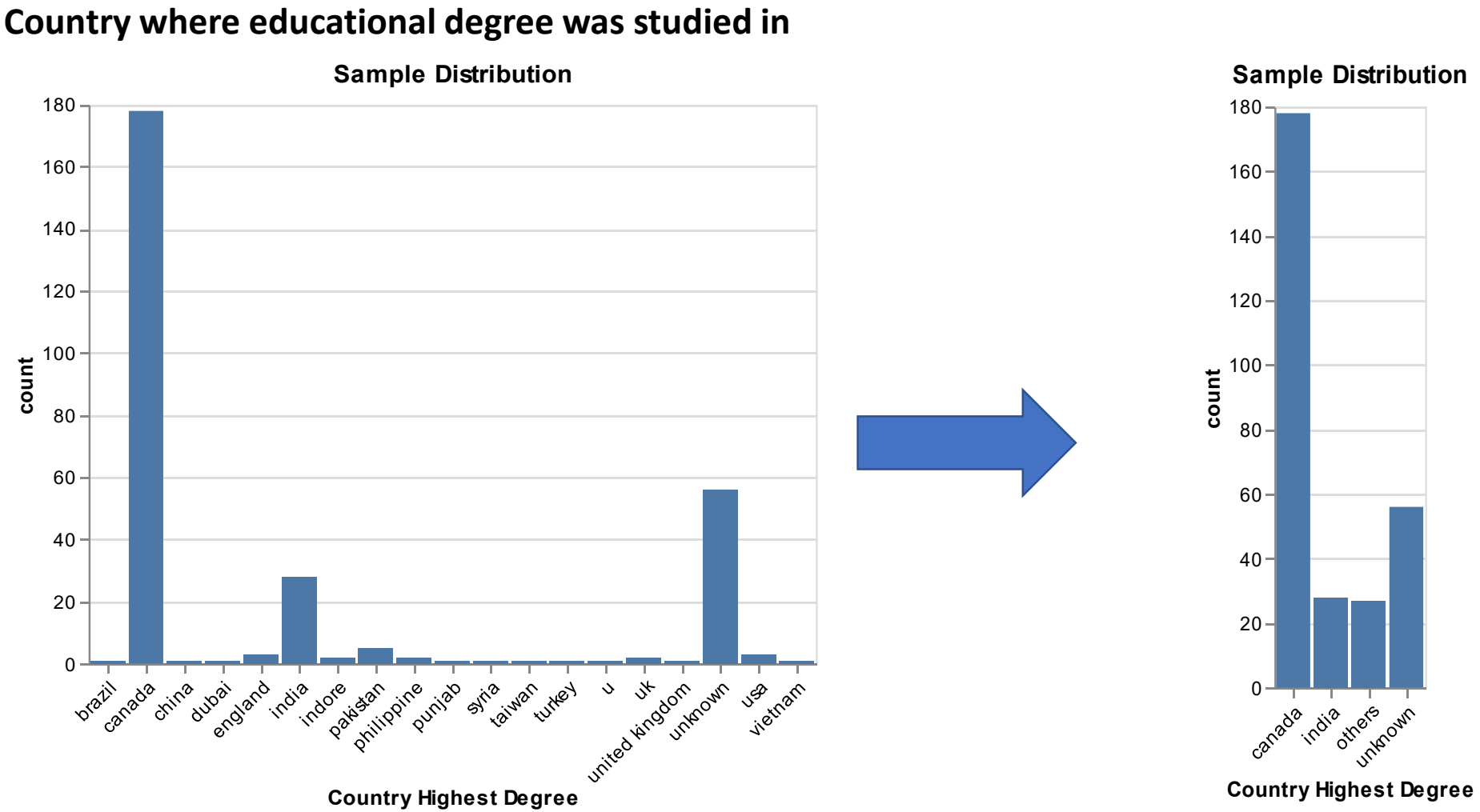
	Neg feats	Neg weights	Pos feats	Pos weights
0	food	-0.265929	sale	0.371016
1	general	-0.216291	mobile	0.346721
2	responsible	-0.176052	lead	0.231738
3	saint	-0.174002	england	0.198113
4	guest	-0.173508	part	0.186862
5	return	-0.161074	vietnam	0.183975
6	sale associate	-0.151548	closing	0.178885
7	production	-0.149070	punjab	0.177460

TF-IDF

BOW and TF-IDF helped us in the feature generation process:

- Number of languages spoken
- Highest educational degree

Feature Engineering - Grouping



- Consolidated low count features to increase interpretability

Methodology – Feature Engineering Summary

N	Feature Category	Number of Features
1	work - experience - position level	15
2	academic background	13
3	knowledge and skills	10
4	work - experience - industry level	7
5	readability - spelling - grammar	3
6	job - tenure - general	5
7	educational level	4
8	job - count	4
9	job - tenure - industry level	3
10	internal Glentel profile	2
11	work - experience - industry level - recency	2
	Grand Total	68

Example Final Modeling Features

knowledge_and_skills
trilingual_flag
volunteer_exp
problem_solving
job_tenure_general
job_hopper
shortest_tenure
education_level
flag_hd_highschool
highest_degree

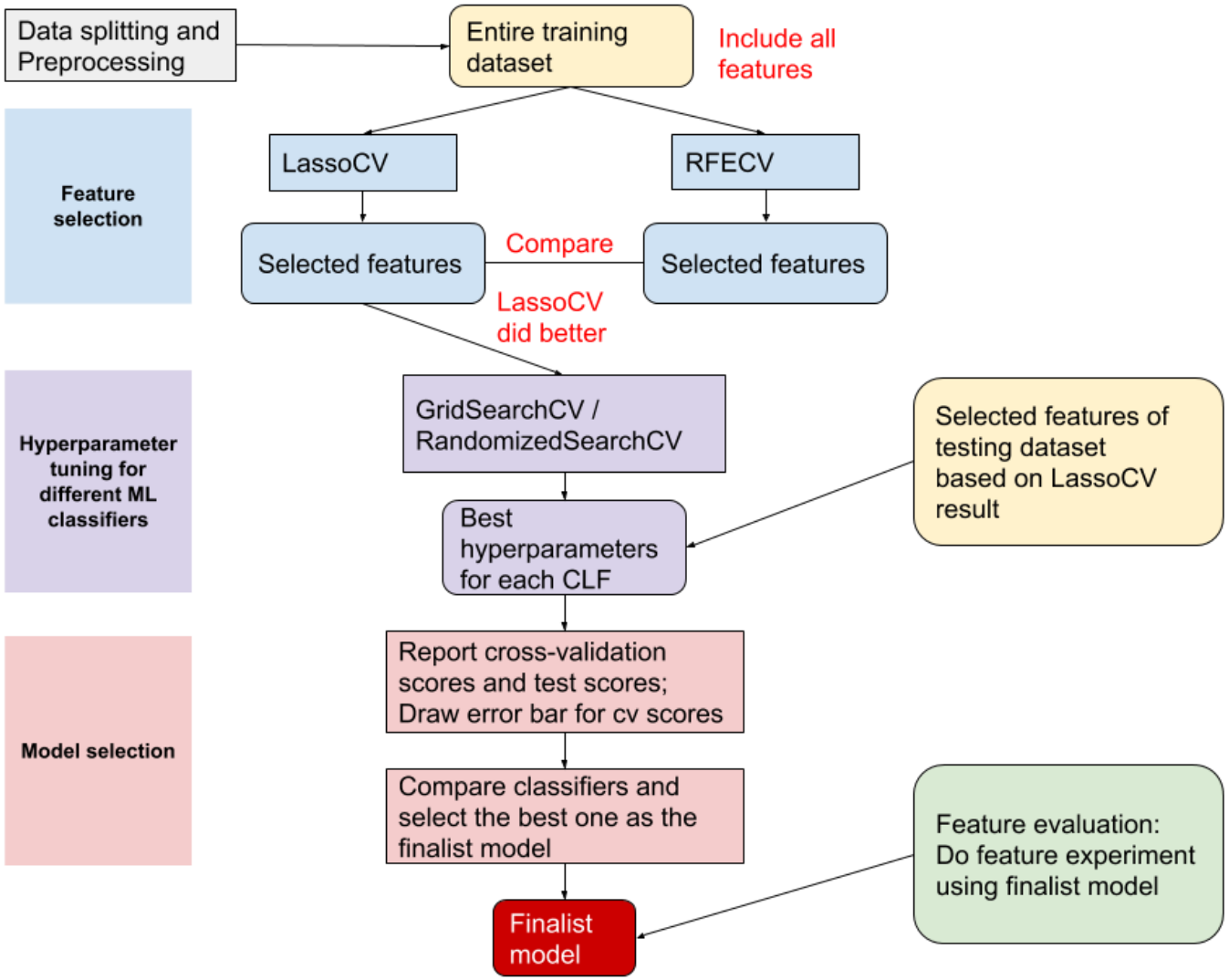
Final Model Set

11 Feature groups

- 61 Numeric features
- 7 Categorical features

Machine Learning Pipeline

Machine Learning Modeling - Workflow



- Data splitting
  - 80/20 train-test
- Preprocessing:
  - One hot encoding, Standardization, Imputation
- Feature selection
  - L1 regularization, RFE
- Hyperparameter tuning
  - Classifiers: Dummy, Logistic Regression, SVM, Random Forest, XGBoost, LGBM, Multi-layer perceptron
  - Handle imbalance data: tune **class_weight**
- Model selection
  - Key metric: **F1 score** (due to imbalance dataset)
    - Cross-validation scores and error bar (many iterations)
    - test scores (final step)
- Feature experiment

Machine Learning Modeling - Feature Selection

LassoCV

Neg feats	Neg weights	Pos feats	Pos weights
flag_hd_highschool	-0.017765	competitor_experience	0.037604
fitness_sports_jobtitle	-0.004068	trilingual_flag	0.031142
accounting_concentration	0.000000	sales_customer_base_exp	0.020959
background_highest_degree_notspecified	0.000000	finance_concentration	0.019978
background_highest_degree_marketing	-0.000000	shortest_tenure	0.009485
background_highest_degree_law	-0.000000	communication_skills	0.004578
background_highest_degree_interactive art technology	-0.000000	cashier_jobtitle	0.001212
background_highest_degree_general	-0.000000	raw_dale_chall_readability	0.000000
background_highest_degree_finance	0.000000	goal_record	0.000000
background_highest_degree_engineering	-0.000000	volunteer_exp	0.000000

- 9 features selected
- All features make sense

REFCV

Neg feats	Neg weights	Pos feats	Pos weights
Food-Convenience-Pharmacy_industry_exp	-1.205933	sports_mention	0.762918
Food_Service_industry_exp	-0.965892	Sport_Travel_Enterntain_Hotel_industry_exp	0.656620
telco_electro_recency	-0.529853	average_tenure_per_job	0.650267
no_lang_spoken	-0.499619	flag_hd_bachelor_plus	0.582243
business_flag	-0.406241	flag_hd_highschool	0.437271
clean_Flesch-Kincaid_readability	-0.389609	general_concentration	0.341969
interactive_arts_and_technology_concentration	-0.286503	no_jobs	0.322376
volunteer_exp	-0.267376	computer_systems_concentration	0.318123
longest_tenure	-0.266800	leadership_mention	0.278352
other_concentration	-0.256938	trilingual_flag	0.271982

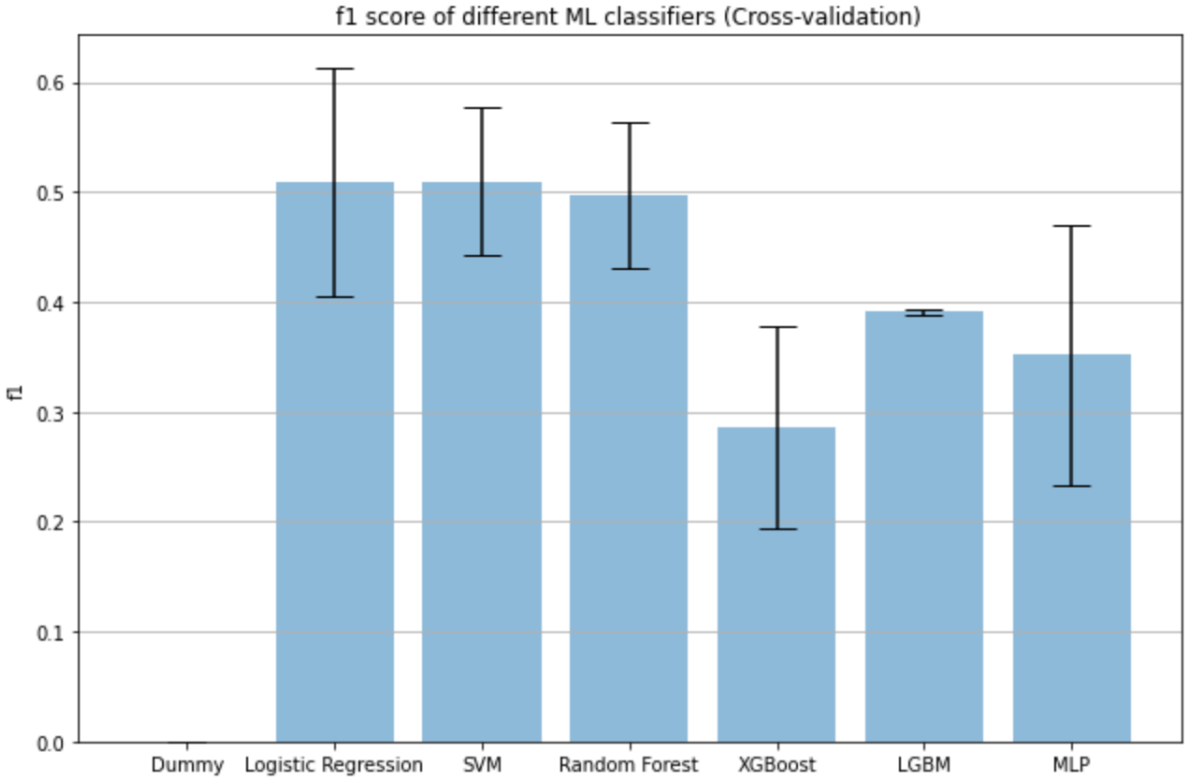
* Only showing 20 features

- 46 features selected
- REF did not handle Multicollinearity well
  - Selected features contradict behavior observed bar plots of features vs HP flag

Use LassoCV selected features

Machine Learning Modeling – Model Selection

F1 Scores with error bar (Cross-validation)



Cross-validation and test scores

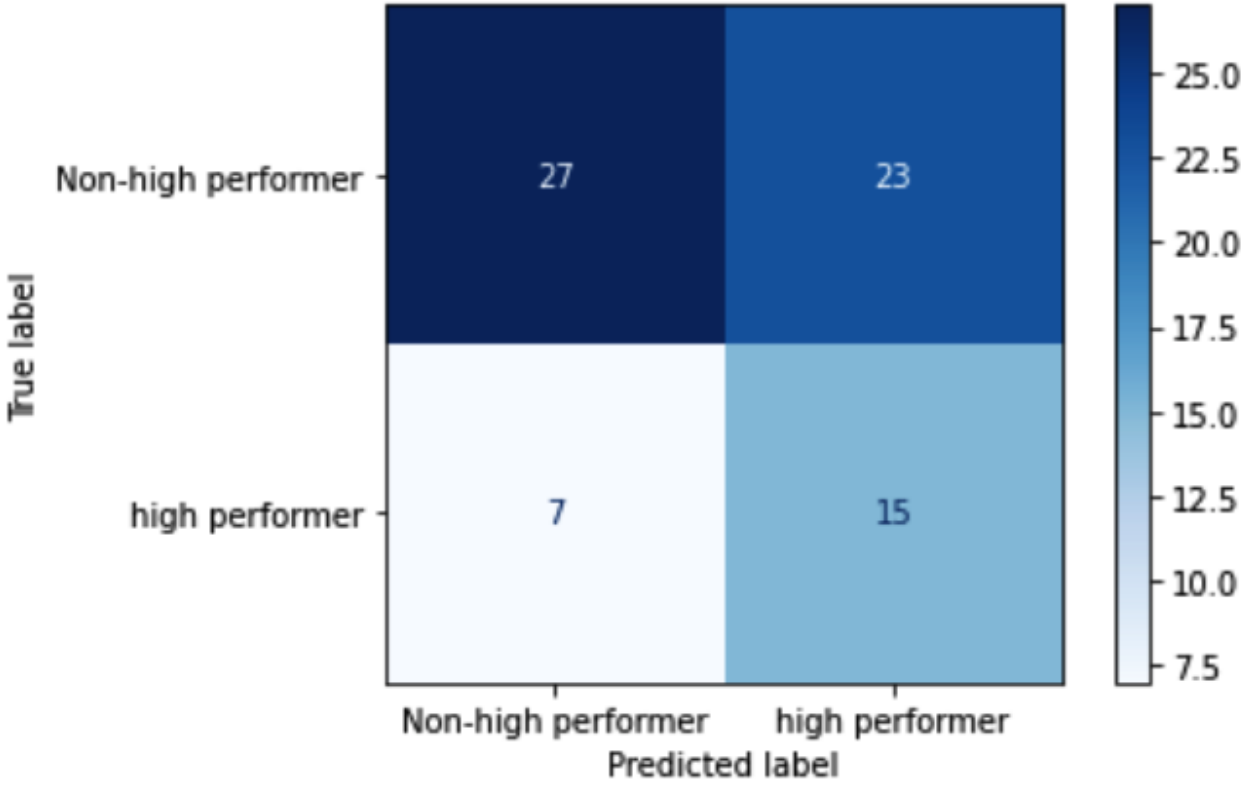
Model	Cross-validation Recall	Cross-validation Precision	Cross-validation F1	test Recall	test Precision	test F1
Dummy	0.000	0.000	0.000	0.000	0.000	0.000
Logistic Regression	0.700	0.402	0.509	0.682	0.395	0.500
SVM	0.643	0.434	0.510	0.636	0.378	0.475
Random Forest	0.614	0.439	0.497	0.318	0.538	0.400
XGBoost	0.229	0.390	0.286	0.182	0.571	0.276
LGBM	1.000	0.243	0.391	0.636	0.333	0.437
Multi-layer Perceptron	0.271	0.594	0.352	0.091	0.400	0.148

Finalist model: **Logistic regression**

better interpretability, less overfitting issue

Machine Learning Modeling – Evaluation

Confusion matrix for high performer in testing dataset



Our finalist model was able to predict ~70% high performers even though only 30% of employees in the testing dataset are high performers

Machine Learning Modeling - Comparison Feature

Features	Cross-validation Recall	Cross-validation Precision	Cross-validation F1	test Recall	test Precision	test F1
Baseline 1: BOW	0.786	0.234	0.359	0.818	0.286	0.424
Baseline 2: TF-IDF	0.971	0.239	0.384	0.591	0.342	0.433
all features	0.700	0.402	0.509	0.682	0.395	0.500
- academic background	0.729	0.377	0.485	0.682	0.429	0.526
- education level	0.714	0.384	0.497	0.591	0.361	0.448
- internal glentel profile	0.700	0.402	0.509	0.682	0.395	0.500
- job counts	0.700	0.402	0.509	0.682	0.395	0.500
- job tenure general	0.700	0.424	0.527	0.591	0.361	0.448
- job tenure industry	0.700	0.402	0.509	0.682	0.395	0.500
- knowledge and skills	0.671	0.284	0.382	0.591	0.419	0.491
- readability	0.700	0.402	0.509	0.682	0.395	0.500
- work experience industry	0.714	0.366	0.478	0.364	0.308	0.333
- work experience industry recency	0.686	0.404	0.498	0.636	0.389	0.483
- work experience position	0.629	0.452	0.522	0.500	0.407	0.449

Baseline:
Use BOW and TF-IDF only

Include all informative features generated by feature engineering

Exclude one group of features each time

Interpretation (Dashboard)

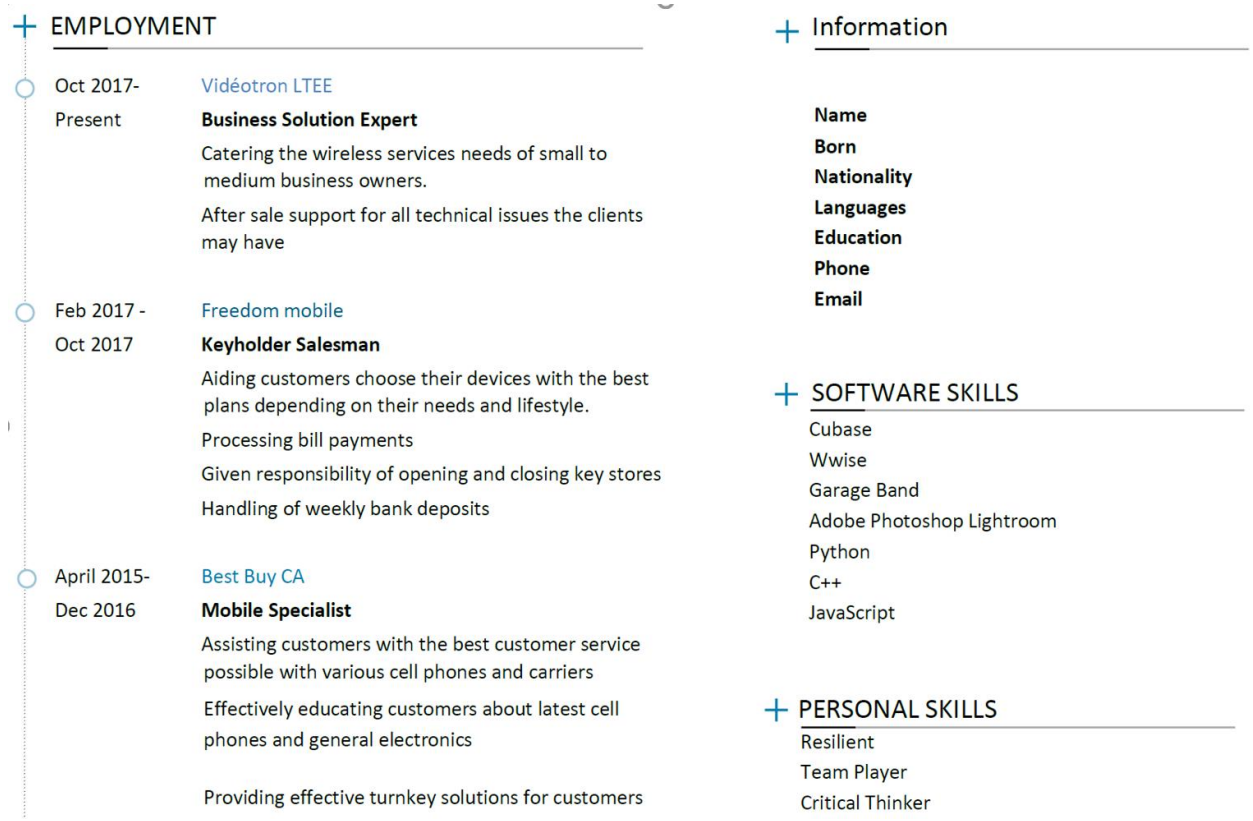
Data Product Handoff

- Documented GitHub repository
- Makefile to reproduce the analysis’s pipeline
- Conda virtual environment with requirements preinstalled

Challenges

- Sample size
- Unstructured resume corpus
- Feature creation and interpretation

Recommendation



EMPLOYMENT

Oct 2017- Present — Vidéotron LTEE
Business Solution Expert
Catering the wireless services needs of small to medium business owners.
After sale support for all technical issues the clients may have

Feb 2017 - Oct 2017 — Freedom mobile
Keyholder Salesman
Aiding customers choose their devices with the best plans depending on their needs and lifestyle.
Processing bill payments
Given responsibility of opening and closing key stores
Handling of weekly bank deposits

April 2015- Dec 2016 — Best Buy CA
Mobile Specialist
Assisting customers with the best customer service possible with various cell phones and carriers
Effectively educating customers about latest cell phones and general electronics
Providing effective turnkey solutions for customers

Information

Name
Born
Nationality
Languages
Education
Phone
Email

SOFTWARE SKILLS

Cubase
Wwise
Garage Band
Adobe Photoshop Lightroom
Python
C++
JavaScript

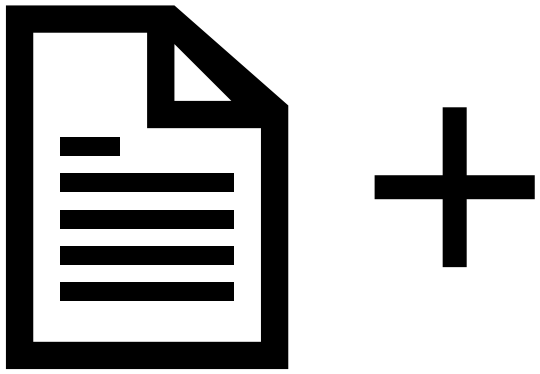
PERSONAL SKILLS

Resilient
Team Player
Critical Thinker

Adobe Photoshop Lightroom
Python
C++
JavaScript
SOFTWARE SKILLS
BestBuy CA
District employee of the month, Aug 2016
Employee of the month, May, June 2016
AWARDS
RESUME
Sept 2018- Music Industry Arts.

Spacy automated information extraction
Automatic annotation tools: Prodigy from Spacy, Inception.

Recommendation



Job 1
- Start Date
- End Date
- Job Title
- Company

Job 2
- Start Date
- End Date
- Job Title
- Company

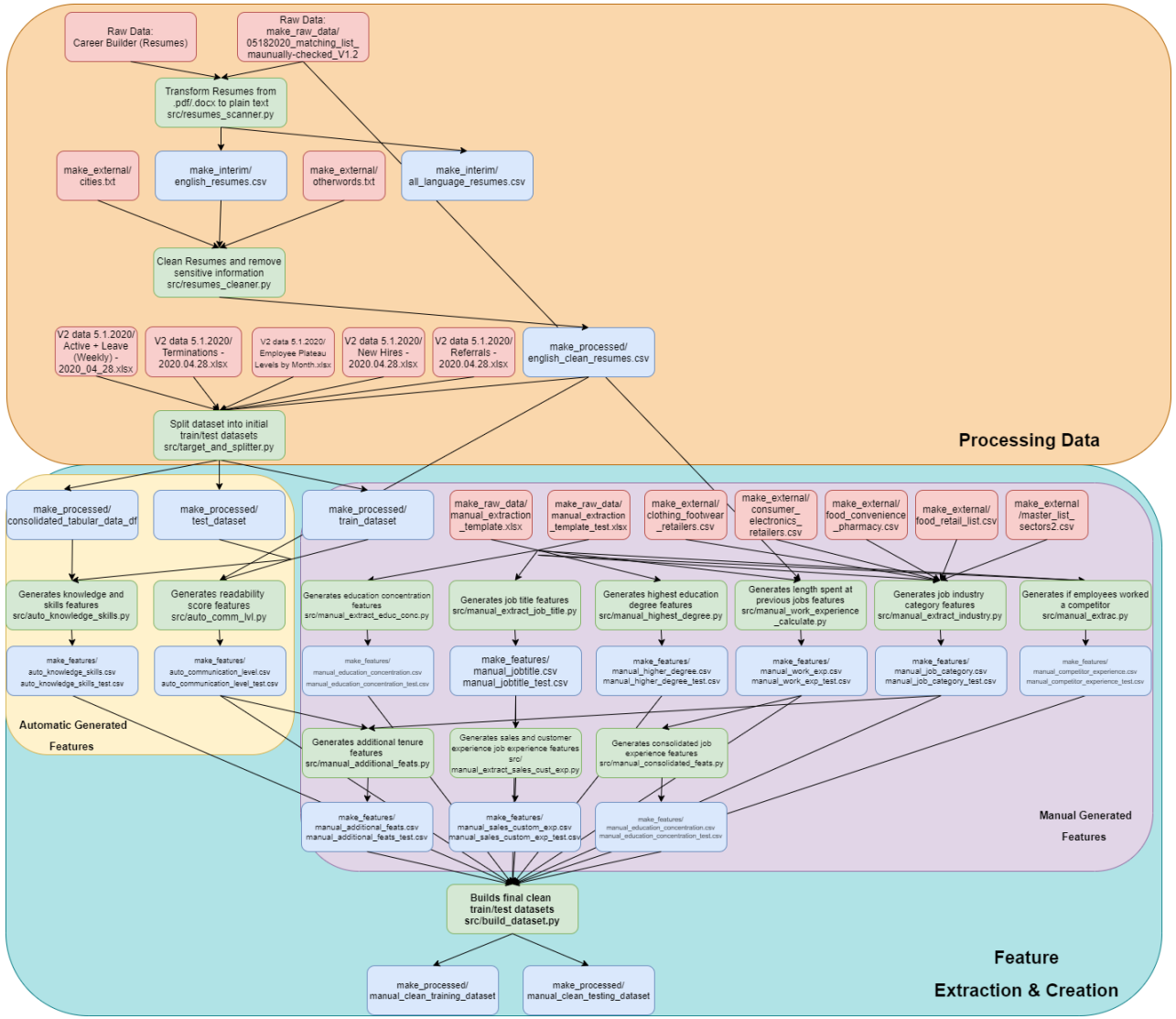
Job 3
- Start Date
- End Date
- Job Title
- Company

Job 4
- Start Date
- End Date
- Job Title
- Company

Questions?

Appendix

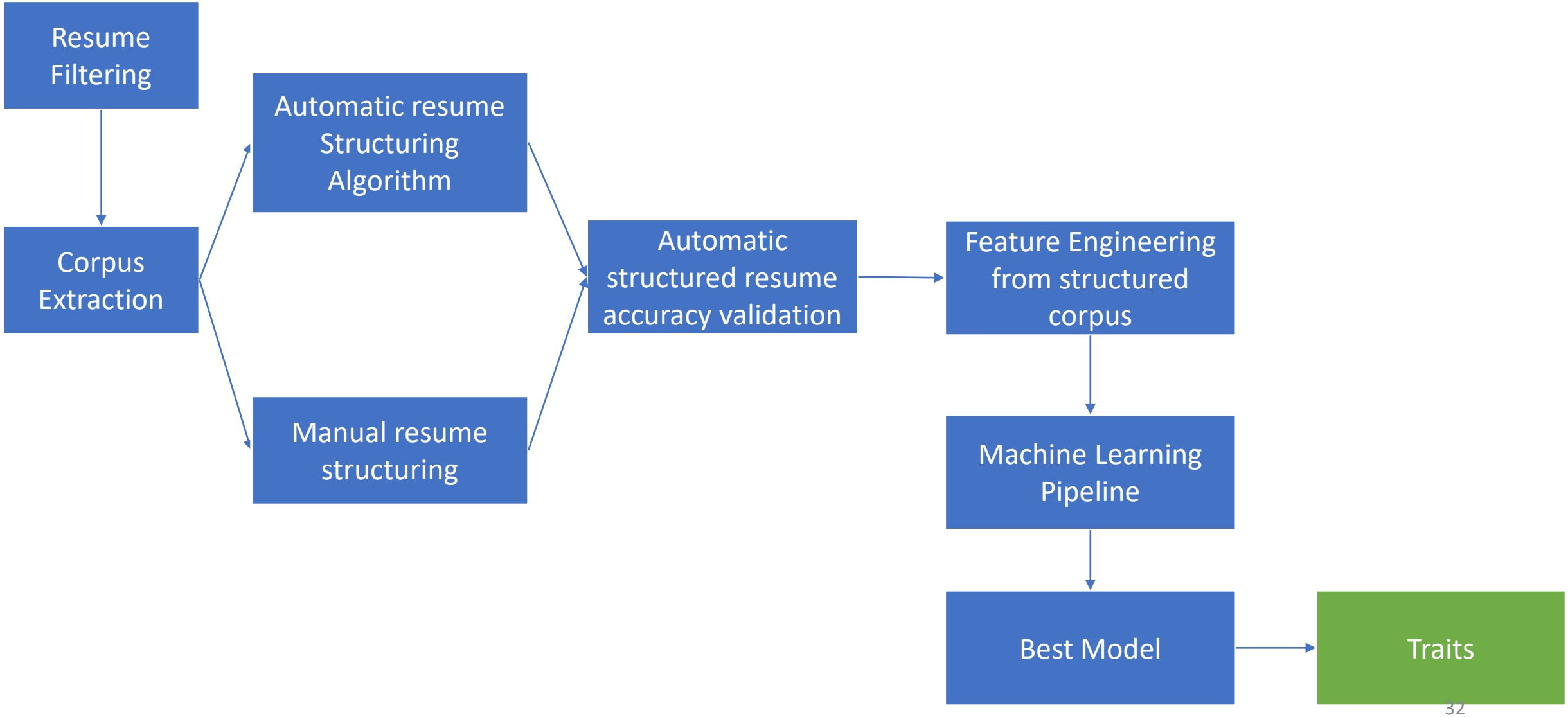
Script Dependency Appendix



P-values of the logistic regression

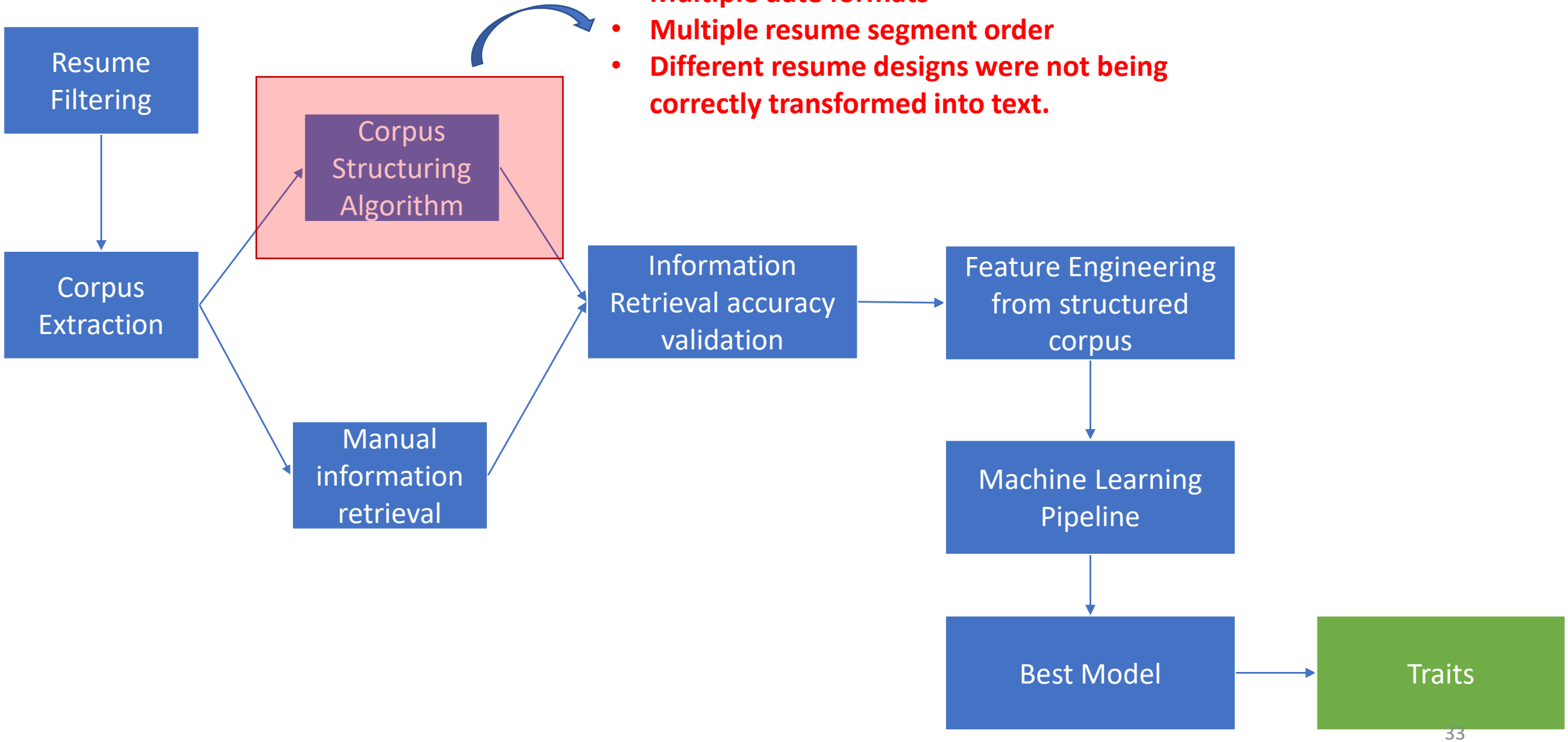
		Coefficients	Standard Errors	t values	p-value	sig
0	(intercept)	-0.2152	0.032	-6.808	0.0	***
1	finance_concentration	0.1903	0.032	5.917	0.0	***
2	flag_hd_highschool	-0.2337	0.033	-7.157	0.0	***
3	shortest_tenure	0.1629	0.032	5.134	0.0	***
4	trilingual_flag	0.2517	0.032	7.783	0.0	***
5	sales_customer_base_exp	0.2995	0.032	9.247	0.0	***
6	communication_skills	0.1308	0.033	3.965	0.0	***
7	competitor_experience	0.3232	0.032	10.055	0.0	***
8	cashier_jobtitle	0.1576	0.032	4.915	0.0	***
9	fitness_sports_jobtitle	-0.2004	0.032	-6.242	0.0	***

Methodology - Expectations



Methodology - Reality

- **Multiple date formats**
- **Multiple resume segment order**
- **Different resume designs were not being correctly transformed into text.**



Information Retrieval

Raw Text/Corpus Extraction

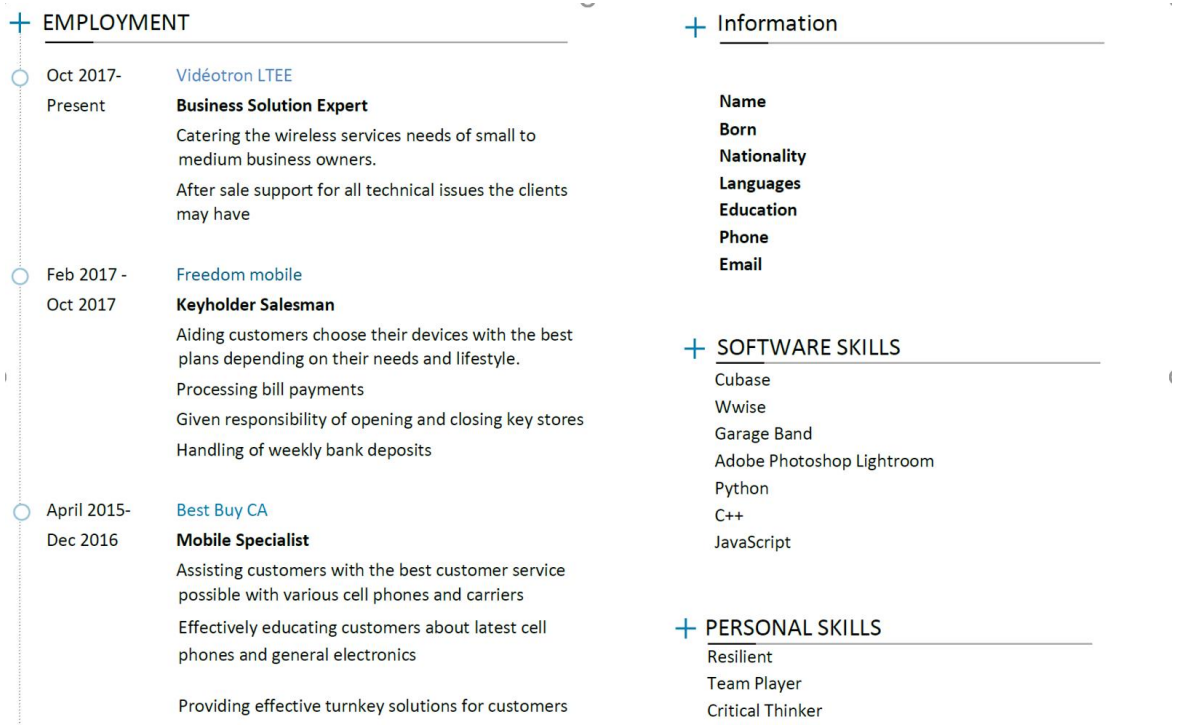
'\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\\n\n\nSKILLS & Competencies\n· Experienced in public relations, rapport building, coordination, collaboration and team-work\n· Strong analytical and mathematical skills\n· Ability to identify issues, propose and implement solutions \n· Excellent verbal and communication skills attained by extensive client dealings \n· Proficient in MS Word, Excel and PowerPoint\n· Strong interpersonal, and communication skills\n· Highly creative and organized\n· Positive energy, “can-do” attitude, adaptive and high degree of initiative\n\n\nWORK EXPERIENCE\n\nTD CANADA TRUST Oct’18-Present \nCustomer Experience advisors.\n\nTHE MOBILE SHOP (Loblaws) Jan’16 – Oct’18 \nSales Associate the cleanliness of cars, gas levels and ensure they were ready for clients\n\n\nEXTRA CURRICULAR ACTIVIES \nUNITED SOUTH ASIANS AT YORK, YORK UNIVERSITY Sept’16 – Apr’16 \nVice President, Finance \n· Ensured \n·Responsible for maintaining a record of the financial standing and to oversee the financial management\n\n\nEDUCATION\nYORK UNIVERSITY Sept’15-Present \nBachelors of Commerce - Financial Economics \n\n\nCERTIFICATIONS\nIFIC (Mutual Funds License) March’18\n \n\n\n'

→

Structured Text

{**'skills'**: ['· experienced in public relations, rapport building, coordination, collaboration and team-work',
'· strong analytical and mathematical skills',
'· ability to identify issues, propose and implement solutions ',
'· excellent verbal and communication skills attained by extensive client dealings ',
'· proficient in ms word, excel and powerpoint',
'· strong interpersonal, and communication skills',
'· highly creative and organized',
'· positive energy, "can-do" attitude, adaptive and high degree of initiative'],
'work experience': ['td canada trust',
'oct’18-present',
'customer experience associate',
'· efficiently handle customers by getting the task done in a timely manner.',
'vice president, finance',
'· ensured budget is adhered according to the approved student association ',
'· responsible for maintaining a record of the financial standing and to oversee the financial management'],
'education': ['york university',
'sept’15- present',
'bachelors of commerce - financial economics',
'certifications',
'ific (mutual funds license)',
'march’18']}

We noticed some resume designs were not being correctly transformed into text:



EMPLOYMENT

Oct 2017- Present — Vidéotron LTEE
Business Solution Expert
Catering the wireless services needs of small to medium business owners.
After sale support for all technical issues the clients may have

Feb 2017 - Oct 2017 — Freedom mobile
Keyholder Salesman
Aiding customers choose their devices with the best plans depending on their needs and lifestyle.
Processing bill payments
Given responsibility of opening and closing key stores
Handling of weekly bank deposits

April 2015- Dec 2016 — Best Buy CA
Mobile Specialist
Assisting customers with the best customer service possible with various cell phones and carriers
Effectively educating customers about latest cell phones and general electronics

Providing effective turnkey solutions for customers

Information

Name
Born
Nationality
Languages
Education
Phone
Email

SOFTWARE SKILLS

Cubase
Wwise
Garage Band
Adobe Photoshop Lightroom
Python
C++
JavaScript

PERSONAL SKILLS

Resilient
Team Player
Critical Thinker

Line order is being altered when transformed into text.

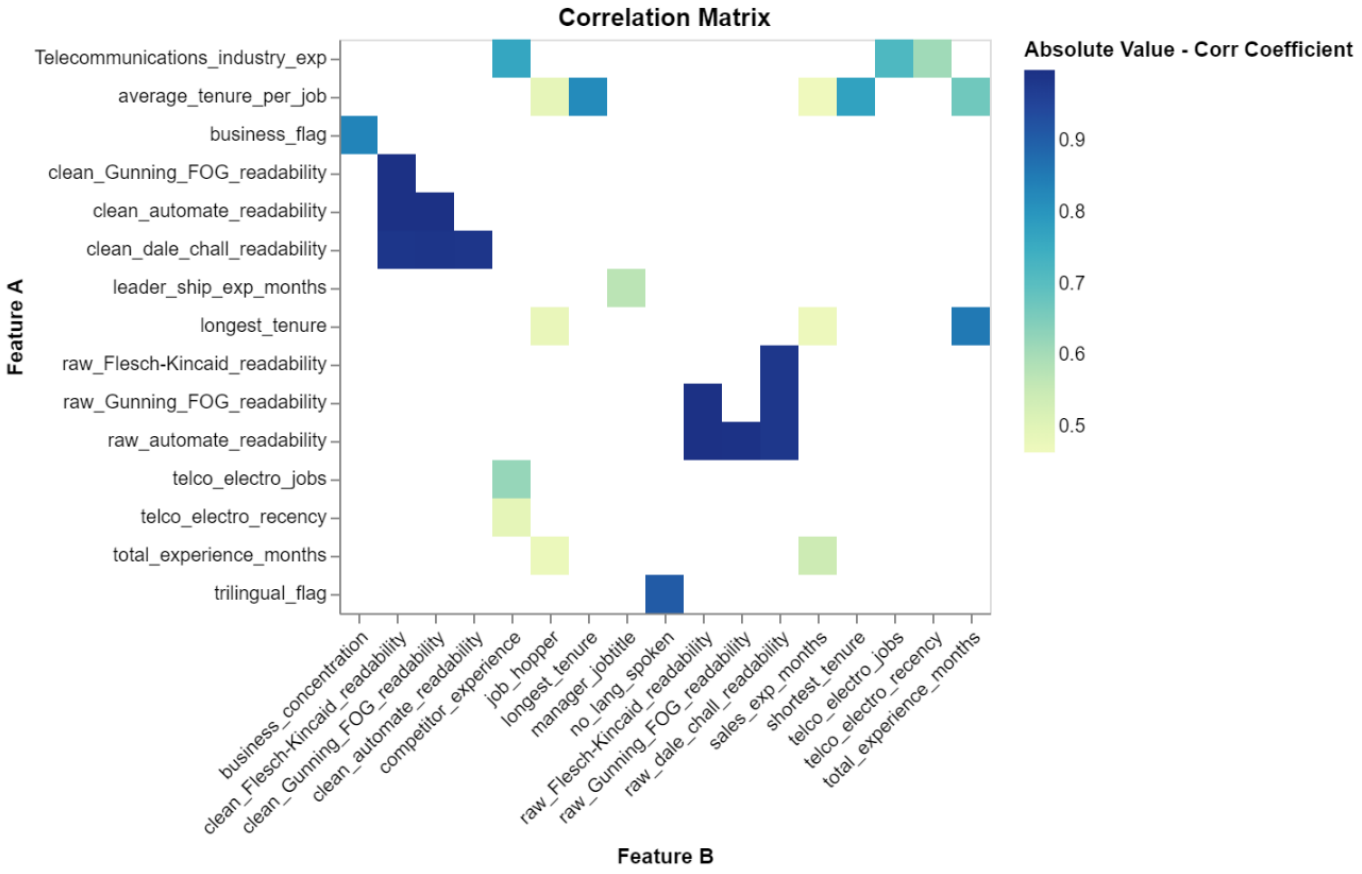
Information
Cubase
Wwise
Garage Band
Adobe Photoshop Lightroom
Python
C++
JavaScript
SOFTWARE SKILLS
BestBuy CA
District employee of the month, Aug 2016
Employee of the month, May, June 2016
AWARDS
RESUME
Sept 2018- Present Music Industry Arts. Dep. Of Art Media and Design Algonquin College
Sept 2013- Algonquin College Bachelor of Computer Science
Sept 2013- April 2016 Bachelor of Computer Science Software Engineering stream Carleton University

We also noticed the lack of structure and consistent patterns within each resume segment.

```
Name: employee_code, dtype: object
                                                                    Out[251]:
['sales associate rogers2018-2019',
 'sales associate  la naturess2015-2017 ',
 'cashier and customer service urban planet 2014-2015 ',
 'cashier (part time) shoppers drug mart 2013-2014',
 'sales representative amarat afghan carpets2010-2013',
 'reference available on request ']
```

```
Name: employee_code, dtype: object
                                                                    Out[254]:
['sales associate/manager in training ',
 'bell canada',
 ' september',
 '2018 - present ',
 '1920 dundas street, london, on n5v 3p1 ',
'shift supervisor/closing manager ',
 'jersey mike's subs',
 ' march 2018 - september 2018 ',
```

Feature Engineering – Correlation Matrix

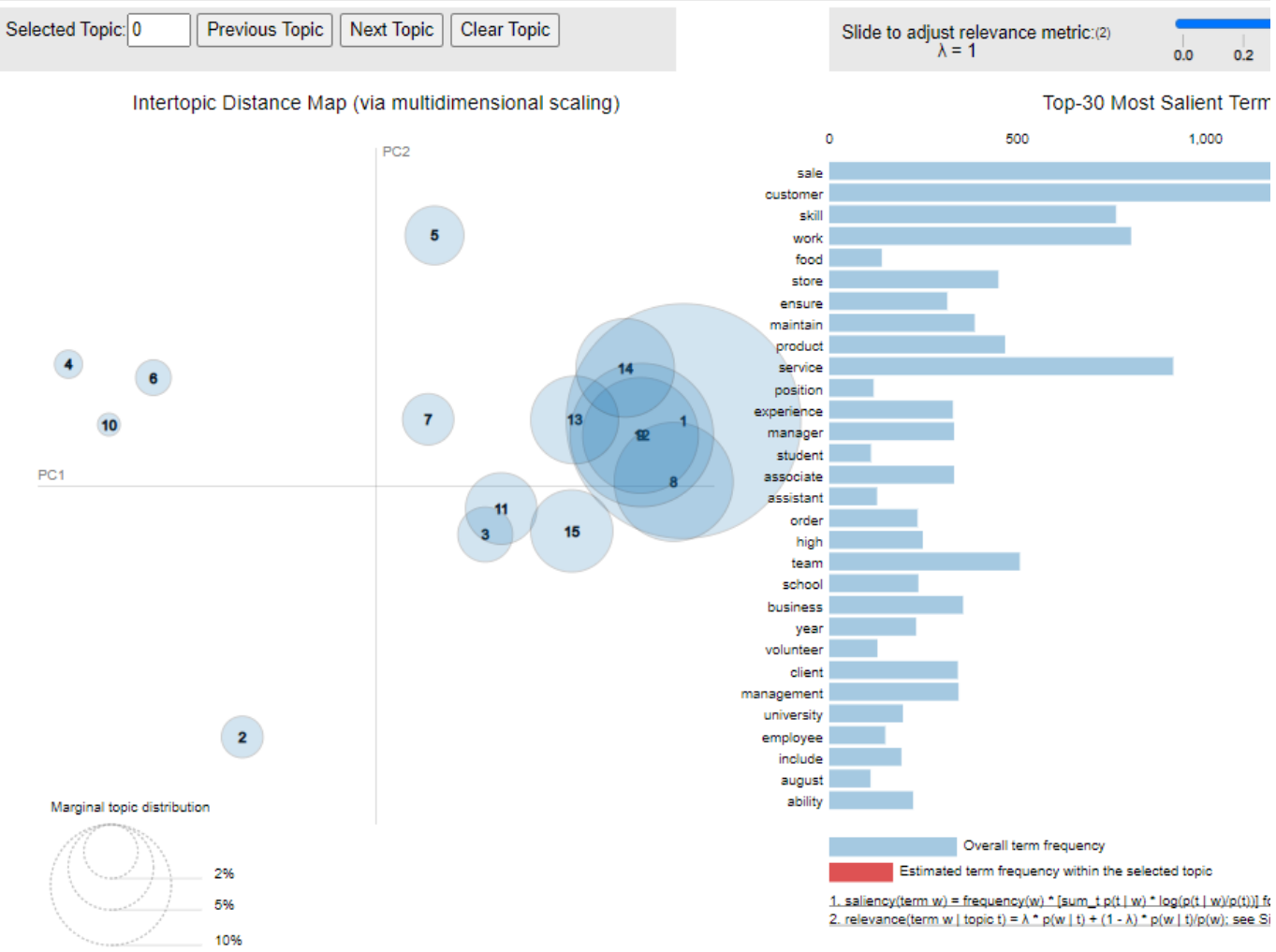


Topic Modelling

All Resumes

```
In [37]:  pyLDAvis.enable_notebook()
          vis = pyLDAvis.gensim.prepare(lda, doc_term_matrix, dictionary, sort_topics=False)
          vis
```

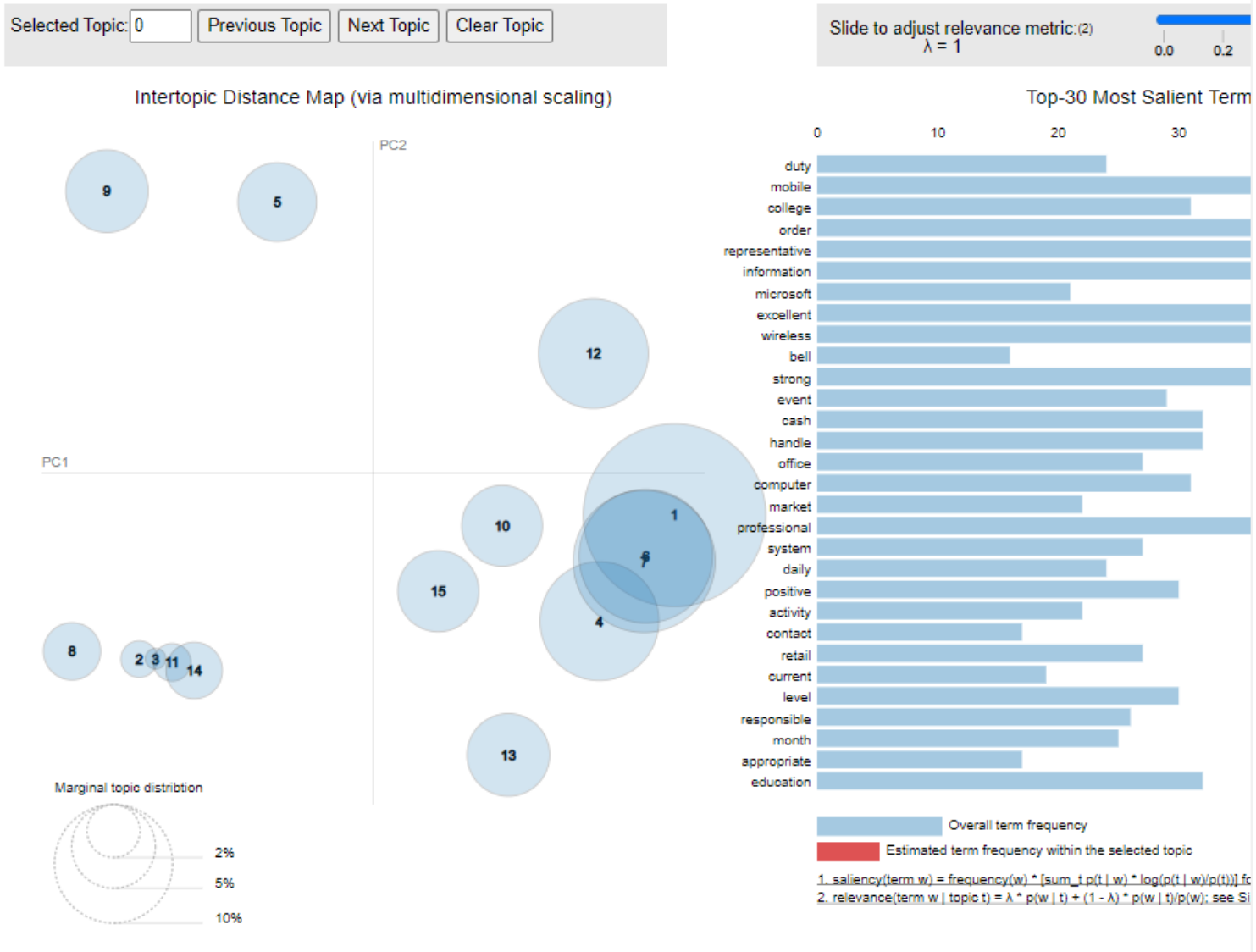
Out[37]:



- Data science
- Fashion
- Nutrition

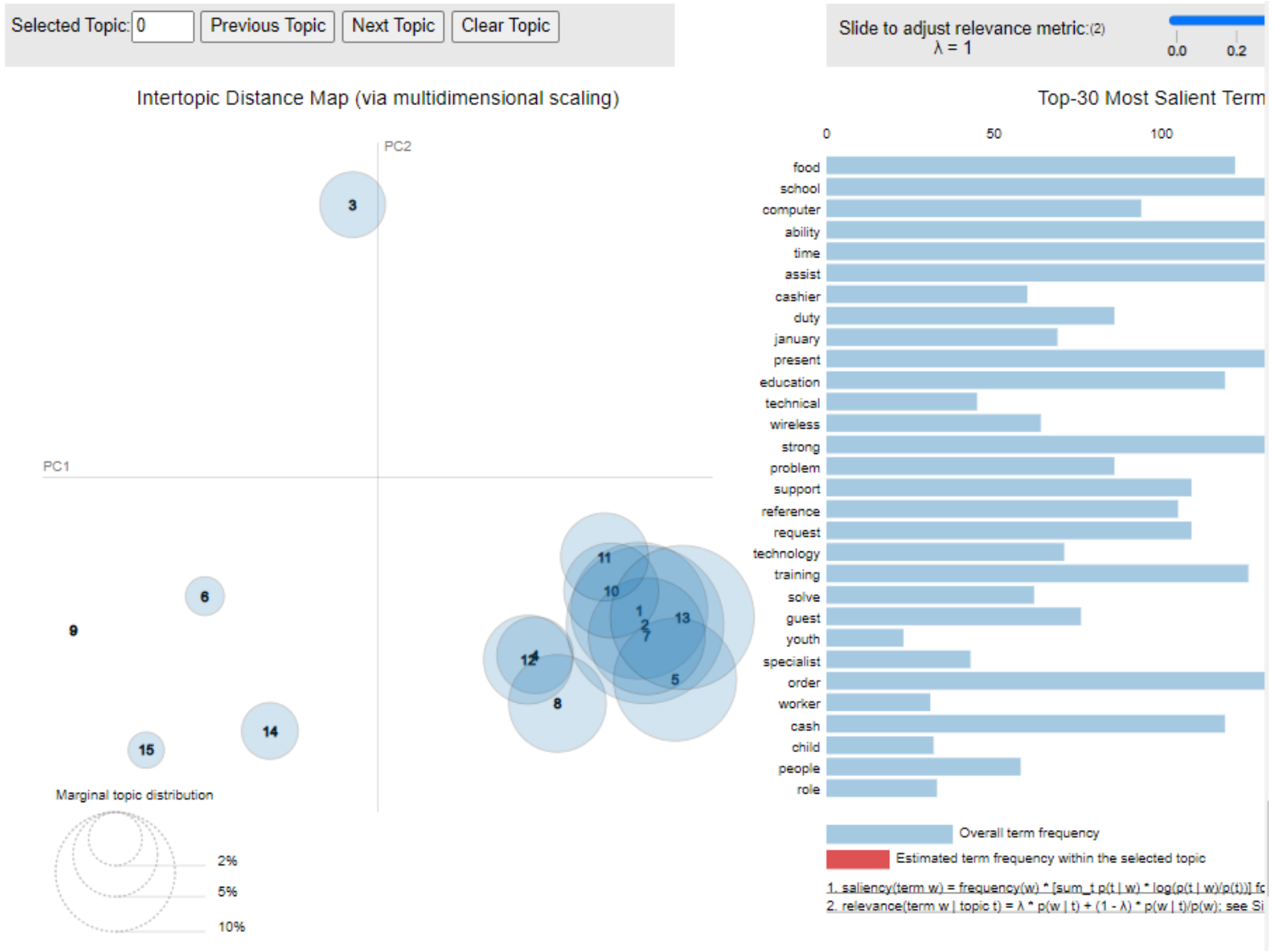
Topic Modelling

High Performers



Topic Modelling

Low Performers



Information Retrieval

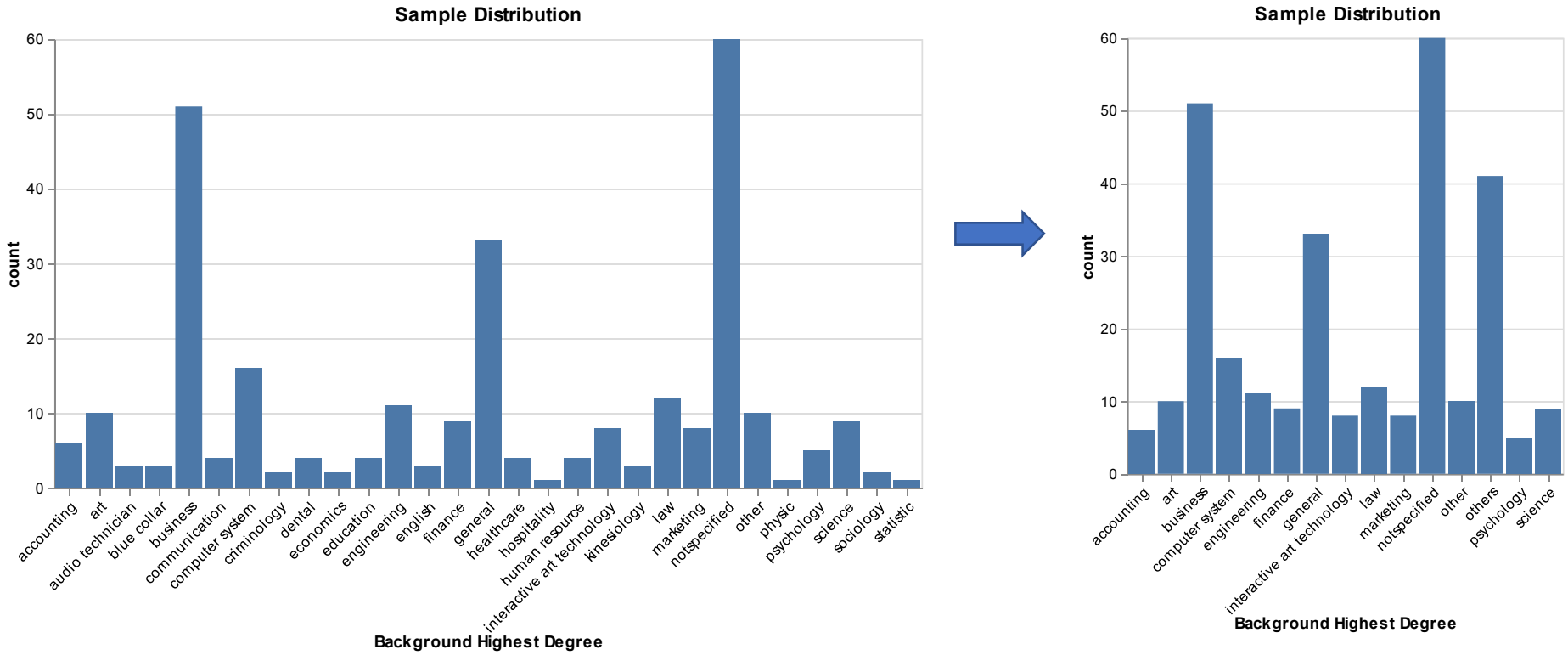
flag	count	perc
not segmented	58	0.2
segmented	230	0.8

Feature Engineering - Features

N	Feature Category	Number of Features
1	work - experience - position level	15
2	academic background	13
3	knowledge and skills	10
4	work - experience - industry level	7
5	readability - spelling - grammar	3
6	job - tenure - general	5
7	educational level	4
8	job - count	4
9	job - tenure - industry level	3
10	internal Glentel profile	2
11	work - experience - industry level - recency	2
	Grand Total	68

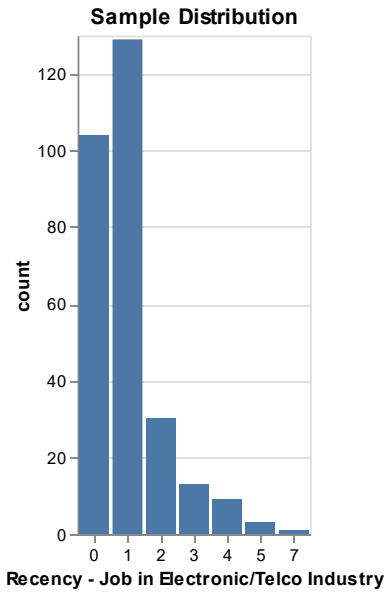
- 68 features
- From 11 feature groups

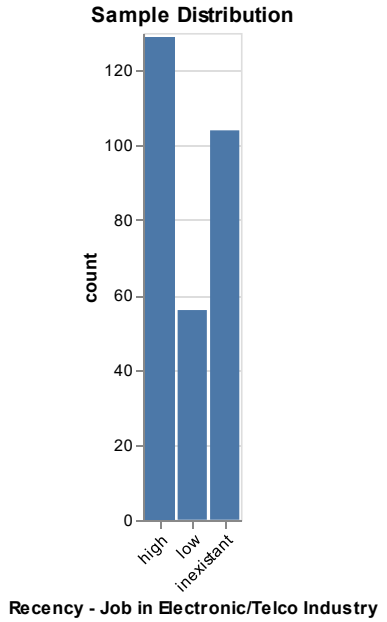
Feature Engineering - Grouping

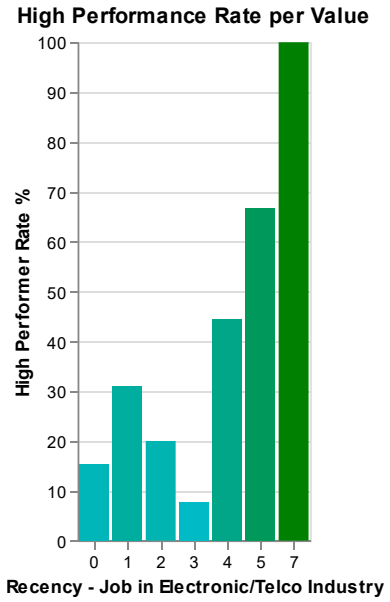


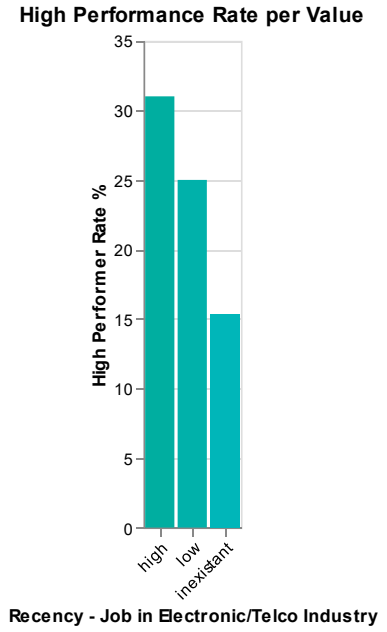
Consolidated low count features to increase interpretability.

Feature Engineering - Grouping









Information Retrieval Example

Automating Competitor Experience

- Started with a difference of 50/146 (1/3) against companies identified manually.
- Many false positive captured for the reference people e-mails
- Had to designed different text cleaning techniques to read data directly and minimize loss of information
- Segmented regex search between key words (Work History, Employment History, volunteered and References)

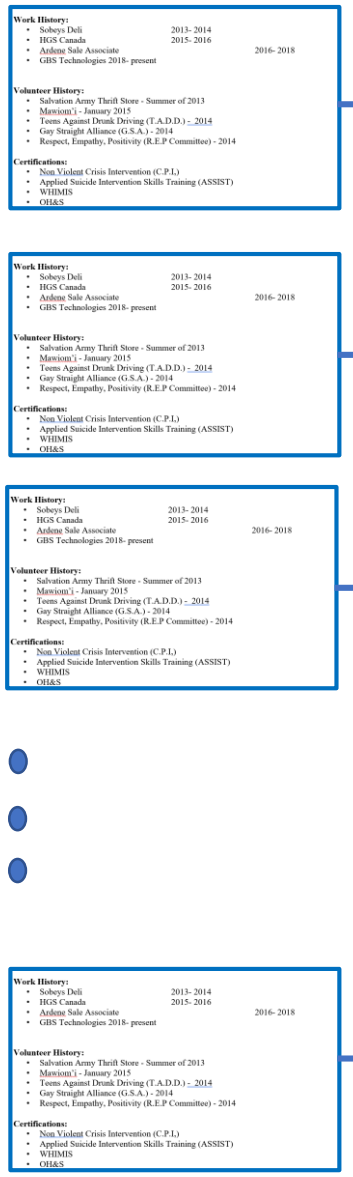
References motioned before competitor job

02/2016 to 10/2017 **Second Assistant Manager**
MCDONALDS – WHISTLER

- Devise and execute strategic recruitment plans aligning with an organization recruitment strategy.
- Validated applicant's references and communicated with previous employers to qualify capabilities and verify work history.
- Followed up with higher management to evaluate employee performance.

- Solving one problem in one CV, many times created a problem in the others.
- Lots of trial and error + lost of trouble shooting time.

```
comp = ['Freedom', 'Koodo', 'Shaw', 'Telus', 'Bell', 'Rogers','The Mobile Shop', 'Best Buy',
        'Videotron', 'Wow[!]* Mobile', 'The Source', 'Walmart', 'Virgin Mobile', 'Osl']
```



Add Basic text cleaner to read from the resume version and not from the clean

```
def clean_txt_f(txt):
    txt = str(txt)
    punct_list = list(string.punctuation+"·"+"–"+"§"+"•"+"—"+"—"+"✓"+"❖"+"❑")
    punct_list.remove("&")
    punct_list.remove("@")
    punct_list.remove("/")
    exclude = set(punct_list)
    clean_txt = "".join([i for i in txt.lower() if i not in exclude])
    clean_txt = re.sub(r'\s+', ' ', clean_txt)
    # Remove Emails
    clean_txt = re.sub(r'([\w\.-]+@[\w\.-]+)', '', clean_txt)
    clean_txt = clean_txt.replace("/", " ")
    return clean_txt
```

```
def find_competitor_list(txt, competitor_lst):
    txt = str(txt).lower()
    competitor_lst = [item.lower() for item in competitor_lst]
    pat_comp =r'.*?(\b(?:{})\b).*?'.format('|'.join(competitor_lst))

    found_list = re.findall(pat_comp, txt)
    # Do not include item found before or after Experience section (Quick & Dirty)
    if ("work experience") in found_list:
        if ("work experience" and "references") in found_list:
            new_list = found_list[found_list.index('work experience'):found_list.index('references')]
        else:
            new_list = found_list[found_list.index('work experience'):]

    elif ("work history") in found_list:
        if ("work history" and "references") in found_list:
            new_list = found_list[found_list.index('work history'):found_list.index('references')]
        else:
            new_list = found_list[found_list.index('work history'):]

    elif ("employment history") in found_list:
        if ("employment history" and "references") in found_list:
            new_list = found_list[found_list.index('employment history'):found_list.index('references')]
        else:
            new_list = found_list[found_list.index('employment history'):]

    elif ("volunteered") in found_list:
        if ("volunteered" and "references") in found_list:
            new_list = found_list[found_list.index('volunteered'):found_list.index('references')]
        else:
            new_list = found_list[found_list.index('volunteered'):]

    elif ("references") in found_list:
        new_list = found_list[:found_list.index('references')]
    else:
        new_list = found_list

    # Remove any reference word used
    ref_list = ['employment history', 'work experience', "work history", 'references', 'volunteered']
    new_list = [x for x in new_list if x not in ref_list]

    return new_list
```

Issue with Spacy for training

- Spacy (Limited extent)
  - Data cleaning
  - Text preprocessing for ML
  - Not for Entity recognition

Spacy Training format template

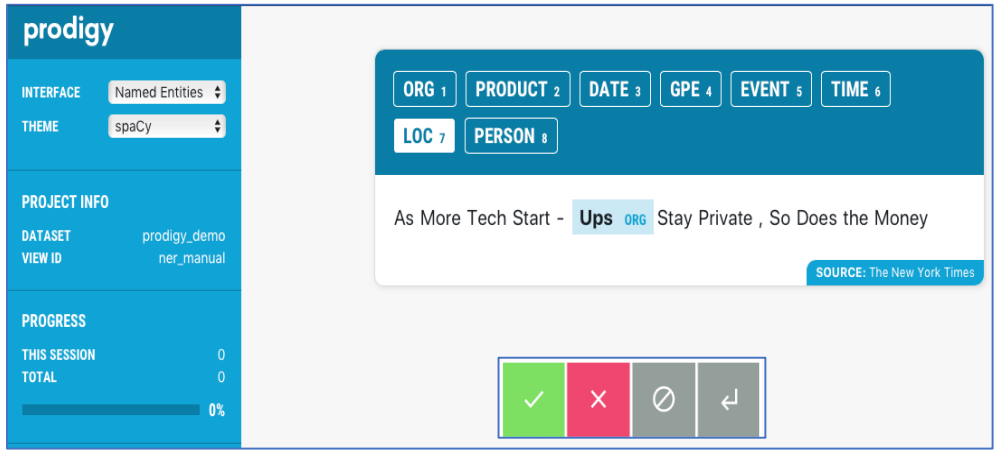
```
train_data = [
    ("Uber blew through $1 million a week", [(0, 4, 'ORG')]),
    ("Android Pay expands to Canada", [(0, 11, 'PRODUCT'), (23, 30, 'GPE')]),
    ("Spotify steps up Asia expansion", [(0, 8, "ORG"), (17, 21, "LOC")]),
    ("Google Maps launches location sharing", [(0, 11, "PRODUCT")]),
    ("Google rebrands its business apps", [(0, 6, "ORG")]),
    ("look what i found on google! 😂", [(21, 27, "PRODUCT")])]
```

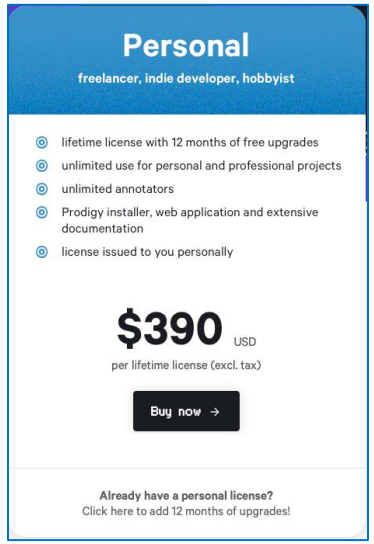
orientation · Strong organizational and time-management abilities · Able to work both independently and in a team environment · Able to work under pressure and multitask when the need arises · Excellent communication skills; Fluent in English, Punjabi, Urdu · Willing to learn and implement new skills and procedures · Diligent and passionate about tasks · Willing to relocate and work in challenging conditions · Energetic, self-motivated and hard-working · Computer aptitude with strong skills in Microsoft office EDUCATION B.A Business Economics September 2014 – July 2018 York University, Faculty of Liberal Arts and Professional Studies, Toronto, ON Diploma in Business Management May 2013 – August 2014 Humber Institute of Technology and Advanced Learning, Toronto, ON WORK & VOLUNTEER EXPERIENCE Wireless Sales Associate June 2018- Present Cellular Point, Etobicoke, ON · Sold wireless services for Koodo, Fido and Virgin, and Lucky Mobile. · Sold electronics and recommended products based on customer needs. · Helped customers solve problems offered solutions and customer service. · Maintained and surpassed sales quota. · Offered technical support to customers. Cashier/Crew Member April 2015 – June 2018 Wendy's, Vaughan, ON · Greet customers as they arrive at the counter · Provide customers with information on add-ons and upsize availability in a bid to upsell food items · Relay orders to the kitchen area and assist in preparing them during slow hours · Tally cash drawers at the end of each shift and ensure that any discrepancies are promptly seen to and resolved · Clean and maintain the counter area and workstations and assist the management with inventory control and stock ordering duties · Listen to and proactively respond to customers' complaints and suggestions with a view to ensure customer loyalty and repeat business Orientation Team Member August 2014 – September 2014 Humber College, Toronto, ON · Greeted and welcomed new students · Assisted new students in registering for orientation sessions and directed them to where they needed to go · Provided relevant orientation information to the new students Auto Parts Manager May 2013 – April 2015 Speed Motor Sports, North York, ON · Maintained stock of auto parts and handling ordering procedures. · Assisted with diagnosing auto problems during busy periods. · Arranged for the ordering of special parts per customer requests. · Inventoried all parts and prepared purchase orders. · Updated database of available parts and adjusted stock accordingly.

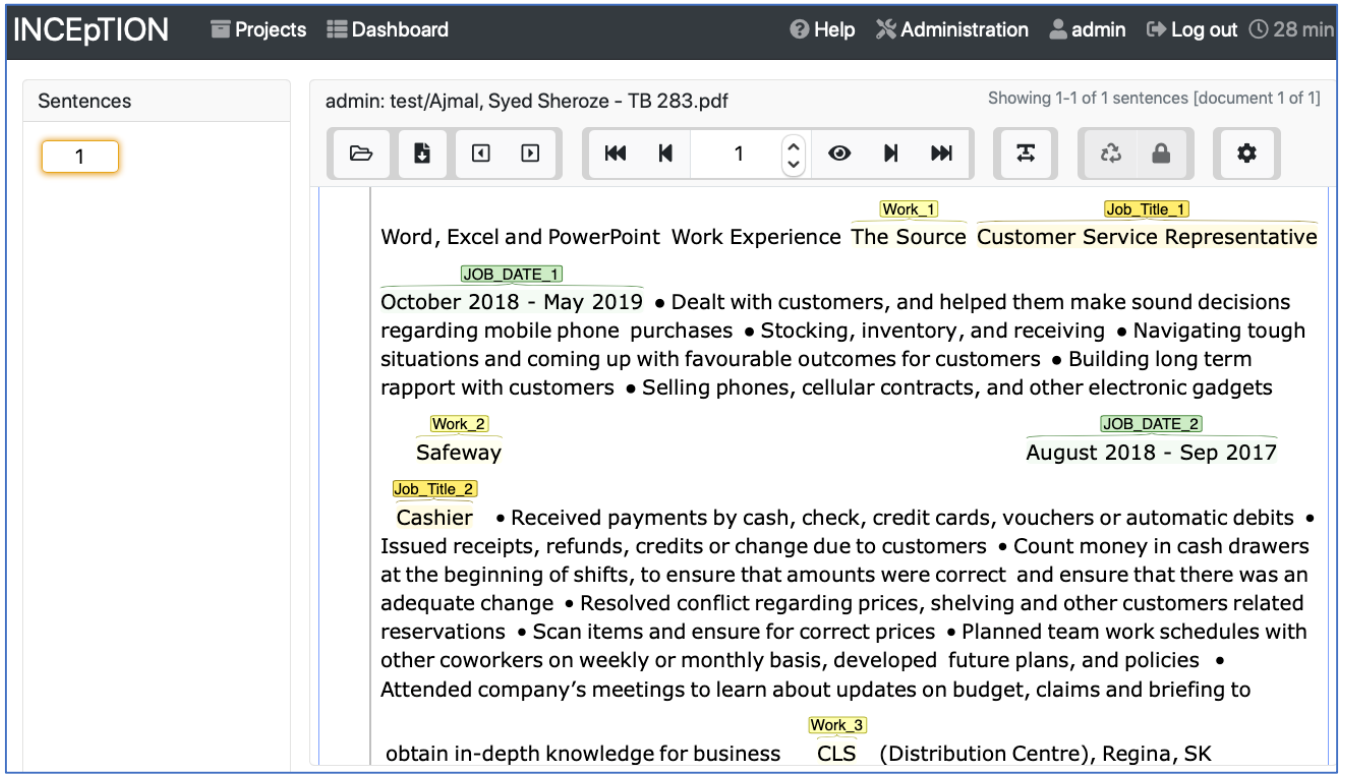
Too many systematic errors would have been generated

Methodology – Spacy + Annotation Tools

"From the Creators of Spacy "

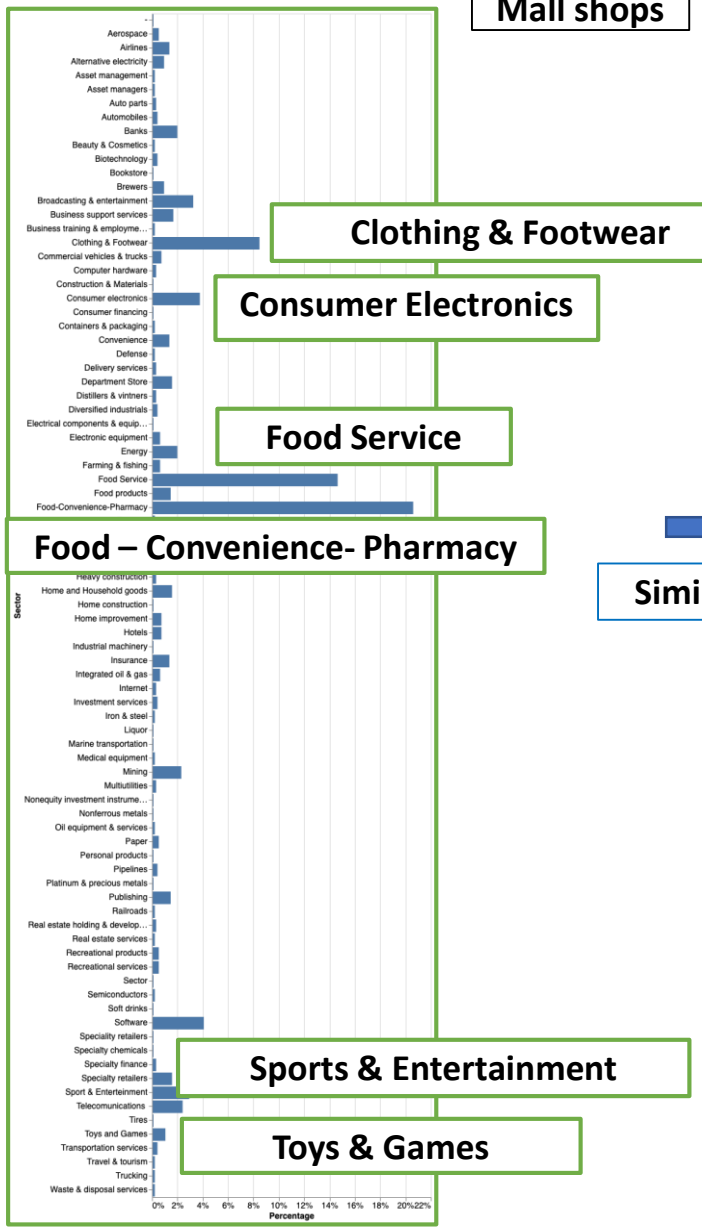






Methodology – Feature Engineering (Industry Characterization)

Wikipedia List



Mall shops

Clothing & Footwear

Consumer Electronics

Food Service

Food – Convenience- Pharmacy

Sports & Entertainment

Toys & Games

Similarity

Levenshtein – distance only

	store_name	perc
industry		
unknown	920	85.264133
Consumer electronics	43	3.985171
Clothing & Footwear	26	2.409639

Key Words

REGEX

- **1079 Companies to identify**
- **15% similarity algorithm.**
- **53% key-words**
- **32% unknown**

Characterized Dataset

industry	store_name	perc
unknown	349	32.344764
Telecomunications	252	23.354958
Food Service	109	10.101946
Sport_Travel_Enterntain_Hotel	65	6.024096
Food-Convenience-Pharmacy	65	6.024096
Clothing & Footwear	59	5.468026
Consumer electronics	29	2.687674
Car Dealers	25	2.316960
Banks	23	2.131603
Beauty & Vanity	19	1.760890
Profesional_Services	19	1.760890
Healthcare	19	1.760890
Blue Collar	15	1.390176
Department Store	11	1.019462
Specialty retailers	6	0.556070
Toys and Games	5	0.463392
Convenience	4	0.370714
Integrated oil & gas	2	0.185357
Home improvement	1	0.092678
Recreational services	1	0.092678
Semiconductors	1	0.092678

Methodology – Feature Engineering Summary

N	Feature Category	Number of Features
1	work - experience - position level	15
2	academic background	13
3	knowledge and skills	10
4	work - experience - industry level	7
5	readability - spelling - grammar	3
6	job - tenure - general	5
7	educational level	4
8	job - count	4
9	job - tenure - industry level	3
10	internal Glentel profile	2
11	work - experience - industry level - recency	2
	Grand Total	68

Final Model Set

- 11 Feature groups
- 61 Numeric features
- 7 Categorical features

68 final feature groups for modeling

	academic_background
1	accounting
2	arts
3	business
4	computer_systems
5	engineering
6	finance
7	general
8	human_resource
9	interactive_arts_and_technology
10	marketing
11	other
12	background_highest_degree ●
13	business_flag

	educational_level
1	highest_degree ●
2	country_highest_degree ●
3	flag_hd_bachelor_plus
4	flag_hd_highschool

	internal_glentel_profile
1	rehired_flag
2	referral_flag

	job_counts
1	telco_electro_jobs
2	no_jobs
3	no_job_categorical ●
4	telco_electro_perc_group ●

	job_tenure_general
1	job_hopper
2	average_tenure_per_job
3	shortest_tenure
4	total_experience_months
5	longest_tenure

	job_tenure_industry
1	sales_exp_months
2	customer_serv_exp_months
3	leader_ship_exp_months

	knowledge_and_skills
1	no_lang_spoken
2	trilingual_flag
3	goal_record
4	sales_customer_base_exp
5	volunteer_exp
6	problem_solver
7	sports_mention
8	communication_skills
9	team_player
10	leadership_mention

	work_experience_industry
1	competitor_experience
2	Clothing_and_Footwear
3	Consumer_electronics
4	Food_Service
5	Food-Convenience-Pharmacy
6	Other_industry
7	Sport_Travel_Enterntain_Hotel

	work_experience_position
1	administrative
2	assistant_manager
3	blue_collar
4	cashier
5	cook
6	customer_service_representative
7	driver
8	education
9	financial_services
10	fitness_sports
11	manager
12	other
13	sales_associate
14	technicians
15	telemarketers

	work_experience_industry_recency
1	recency_type_telco_electro_exp ●
2	telco_electro_recency

	readability
1	raw_dale_chall_readability
2	clean_Flesch-Kincaid_readability
3	read_score_categorical ●

	Word Vectorizer
1	Bag of Words
2	Tf-idf

● Categorical features

Appendix

	cross-validation					test				
	accuracy	recall	precision	roc_auc	f1	accuracy	recall	precision	roc_auc	f1
model										
Dummy	0.757	0.000	0.000	0.500	0.000	0.694	0.000	0.000	0.500	0.000
Logistic Regression	0.670	0.700	0.402	0.744	0.509	0.583	0.682	0.395	0.585	0.500
SVM	0.698	0.643	0.434	0.708	0.510	0.569	0.636	0.378	0.590	0.475
Random Forest	0.695	0.614	0.439	0.680	0.497	0.708	0.318	0.538	0.582	0.400
XGBoost	0.726	0.229	0.390	0.642	0.286	0.708	0.182	0.571	0.660	0.276
LGBM	0.243	1.000	0.243	0.500	0.391	0.500	0.636	0.333	0.626	0.437
Malti-layer Perceptron	0.774	0.271	0.594	0.693	0.352	0.681	0.091	0.400	0.590	0.148

Appendix

	cross-validation					test				
	accuracy	**recall**	**precision**	**roc_auc**	**f1**	**accuracy**	**recall**	**precision**	**roc_auc**	**f1**
Baseline 1: bag-of-word with lr	0.340	0.786	0.234	0.505	0.359	0.319	0.818	0.286	0.509	0.424
Baseline 2: tf-idf with lr	0.243	0.971	0.239	0.540	0.384	0.528	0.591	0.342	0.563	0.433
all features	0.670	0.700	0.402	0.744	0.509	0.583	0.682	0.395	0.585	0.500
- academic_background	0.621	0.729	0.377	0.735	0.485	0.625	0.682	0.429	0.607	0.526
- educational_level	0.649	0.714	0.384	0.718	0.497	0.556	0.591	0.361	0.576	0.448
- internal_glentel_profile	0.670	0.700	0.402	0.744	0.509	0.583	0.682	0.395	0.585	0.500
- job_counts	0.670	0.700	0.402	0.744	0.509	0.583	0.682	0.395	0.585	0.500
- job_tenure_general	0.688	0.700	0.424	0.729	0.527	0.556	0.591	0.361	0.583	0.448
- job_tenure_industry	0.670	0.700	0.402	0.744	0.509	0.583	0.682	0.395	0.585	0.500
- knowledge_and_skills	0.541	0.671	0.284	0.693	0.382	0.625	0.591	0.419	0.634	0.491
- readability	0.670	0.700	0.402	0.744	0.509	0.583	0.682	0.395	0.585	0.500
- work_experience_industry	0.621	0.714	0.366	0.729	0.478	0.556	0.364	0.308	0.558	0.333
- work_experience_industry_recency	0.659	0.686	0.404	0.739	0.498	0.583	0.636	0.389	0.578	0.483
- work_experience_position	0.719	0.629	0.452	0.719	0.522	0.625	0.500	0.407	0.573	0.449

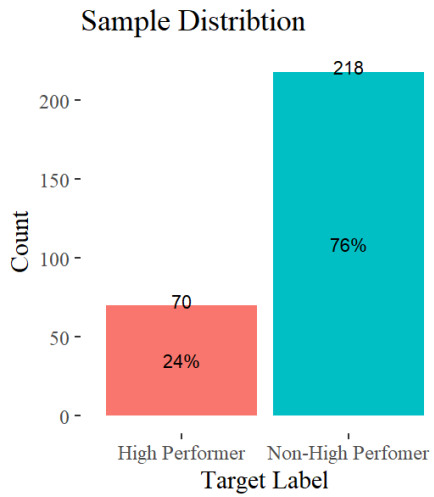
Traits of high performers in Glentel

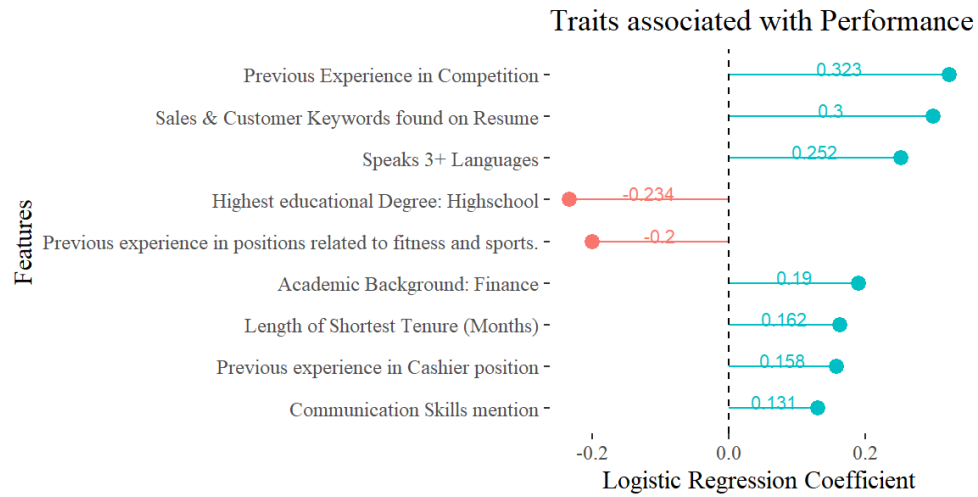
Features

- (•) Highest educational Degree: Highschool
- ( ) Academic Background: Finance
- ( ) Previous Experience in Competition
- ( ) Previous experience in Cashier position
- ( ) Previous experience in positions related to fitness and sports.
- ( ) Sales & Customer Keywords found on Resume
- ( ) Length of Shortest Tenure (Months)
- ( ) Speaks 3+ Languages
- ( ) Communication Skills mention

Overview | Individual Feats

Overall Target Distribution





Traits of high performers in Glentel

Features

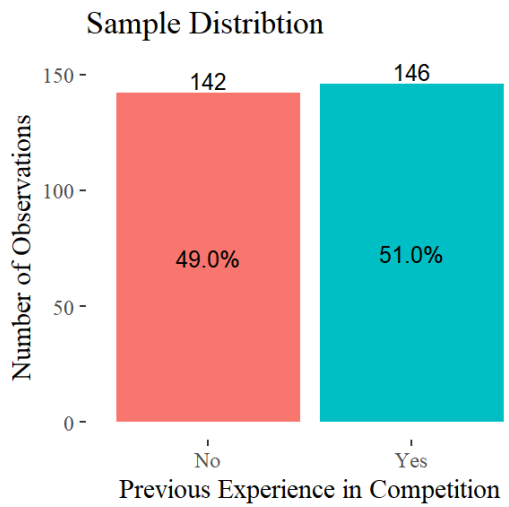
- Highest educational Degree: Highschool
- Academic Background: Finance
- Previous Experience in Competition
- Previous experience in Cashier position
- Previous experience in positions related to fitness and sports.
- Sales & Customer Keywords found on Resume
- Length of Shortest Tenure (Months)
- Speaks 3+ Languages
- Communication Skills mention

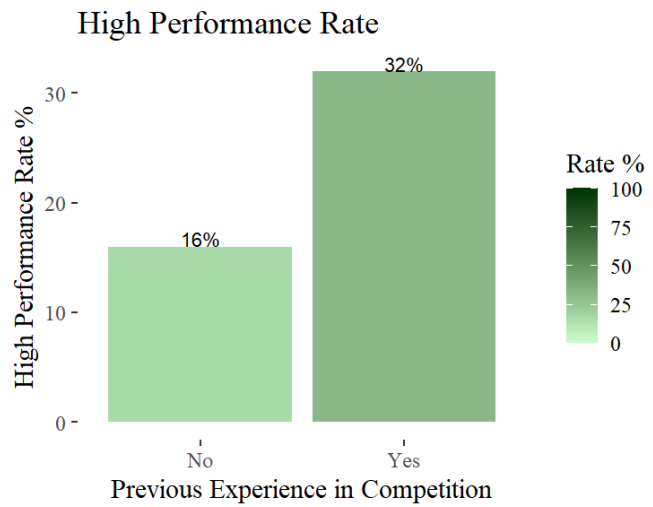
Overview | Individual Feats

Previous Experience in Competition

Indicates if applicant has previously worked for companies considered as the competition for Glentel.
This includes the following companies: Freedom, Koodo, Shaw, Telus, Bell, Rogers, The Mobile Shop, Best Buy, Videotron, WOW! Mobile, The Source, Walmart,Virgin, OSL

Applicants who habe previuosly worked with these companies have shown a higher proportion of high performers.





Traits of high performers in Glentel

Features

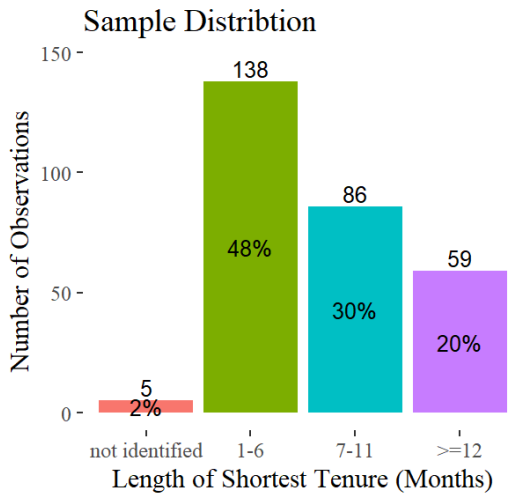
- Highest educational Degree: Highschool
- Academic Background: Finance
- Previous Experience in Competition
- Previous experience in Cashier position
- Previous experience in positions related to fitness and sports.
- Sales & Customer Keywords found on Resume
- Length of Shortest Tenure (Months)
- Speaks 3+ Languages
- Communication Skills mention

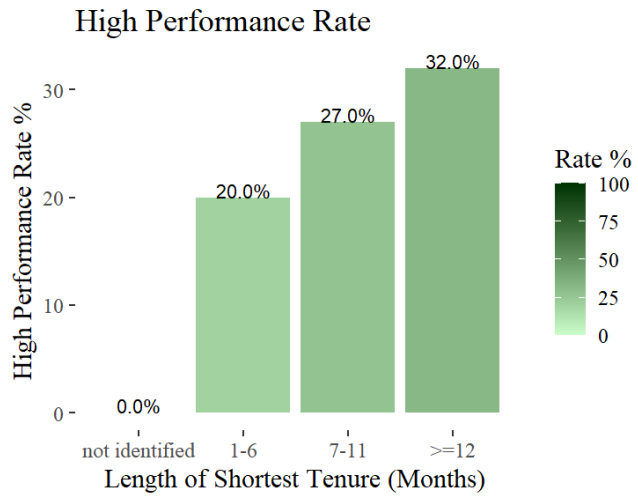
Overview | Individual Feats

Length of Shortest Tenure (Months)

Indicates if the length in months of the applicant's shortes tenure.

The longer the applicant's shortes tenure the higher the proportion of high performers.





Traits of high performers in Glentel

Features

- Highest educational Degree: Highschool
- Academic Background: Finance
- Previous Experience in Competition
- Previous experience in Cashier position
- Previous experience in positions related to fitness and sports.
- Sales & Customer Keywords found on Resume
- Length of Shortest Tenure (Months)
- Speaks 3+ Languages
- Communication Skills mention

Overview | Individual Feats

Speaks 3+ Languages

Indicates if applicant has indicated the speaking of 3 or more languages in the resume.
This was identified by searching for language names in the applicants resume.

Languagues looked for in the resume are: 'spanish', 'french', 'german', 'punjabi', 'hindi', 'urdu', 'arabic', 'mandarin', 'dari', 'japanese', 'filipino', 'tamil', 'cantonese', 'russian'. All applicant are assumed to speak english.

Applicants that speak 3 or more languages have shown a greater proportion of high-performers.