



Indeed Job Postings **Skill Match**

DATS 6202 - Machine Learning I, Final Project Report

Brian Wilson, Joseph Francis, Rashmi Menon, Srilatha Lakka, Renee Adonteng

Introduction

Problem Statement

The positive headlines heralding constant U.S. economic growth since the 2008 financial crisis mask the dynamics of the US economy and the geographically-uneven distribution of that growth. The largest metropolitan regions of the US accounted for “two-thirds of output growth on the economic front and 73 percent of employment gains between 2010 and 2016—figures that actually have increased since 2014, when they reached nearly 72 and 74 percent. By contrast, smaller metropolitan areas with less than 250,000 people—representing 9 percent of the nation’s population—have lost ground. Since 2010, in fact, these communities made a negative contribution of -6.5 percent to the nation’s growth, with their contribution ticking up modestly in the last two years and their output and employment growth contributions declining to less than 3 percent and 5 percent of the national total, respectively. As for the rural tier, the trends have been even worse. By the 2014 to 2016 period, rural communities’ contribution to national population growth had turned negative and the ebbing of the earlier oil and gas boom saw output and employment growth decline precipitously as a share of national gains.”¹

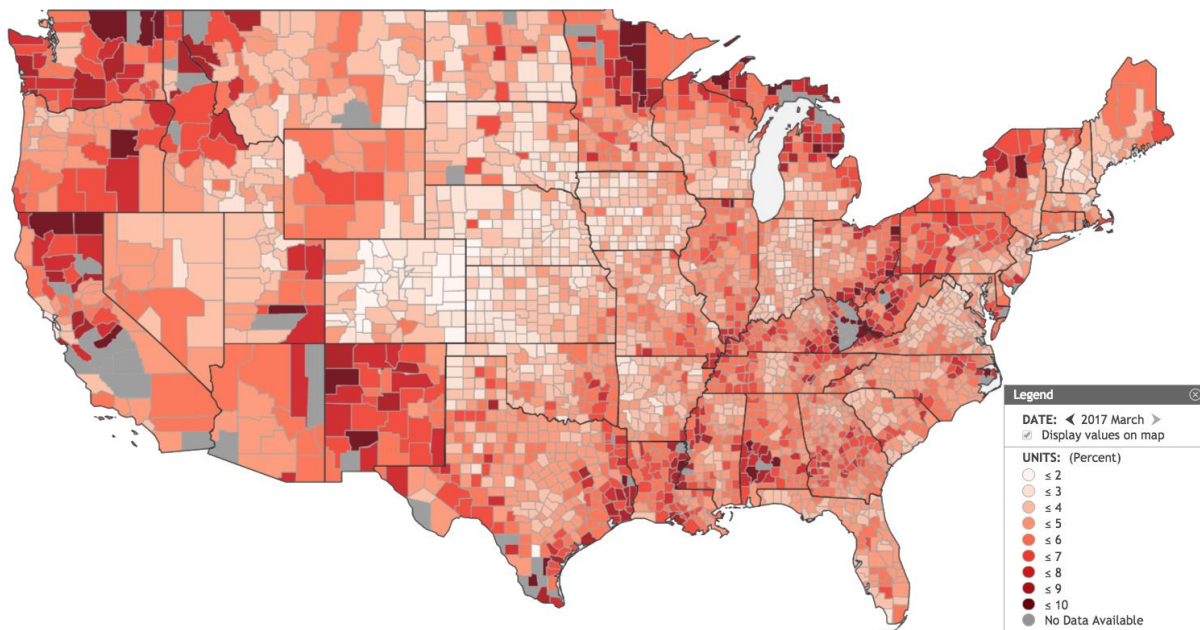


Figure 1: Unemployment rate by county in the United States (March 2017) ²

¹ <https://www.brookings.edu/blog/the-avenue/2018/01/22/uneven-growth/>

² https://en.wikipedia.org/wiki/Unemployment_in_the_United_States#/media/File:Unemployment_by_county_in_the_United_States.png

The dislocated workforce in these economically-depressed areas of the US require job training assistance to develop marketable skills. The diversity and geographic sprawl of the US economy prevents a one-size fits-all solution and thus demands programs and services tailored to each region. The U.S. Department of Labor's Employment and Training Administration (ETA), which allocates federal funds for a collection of training programs and services for dislocated workers³, as well as their equivalent organizations on the state and local governmental-level, must identify and target through training the skills required by industry in each geographic area they support. The impacted workers themselves should pursue skills currently in-demand by industry near them or consider relocation to geographic areas offering opportunities to apply their current skillset.

Business Objectives

This paper will propose applying machine learning and natural language processing techniques to extract skills, educational requirements, experience qualifications, and salary details from job postings across industries throughout the US. Furthermore, it will propose a method by which each job posting is "geocoded", thus providing a location context to those extracted attributes. The data warehouse built from this extracted and geocoded data, would provide a real-time view into the requirements and needs of industry across the US.

A business intelligence dashboard built upon the final data set would allow governmental, educational, and non-profit agencies focused on workforce retraining to explore each geographic region's needs and inform their curriculum. Dislocated workers could plan the next phase of their careers, set salary expectations based upon industry, location, and experience, as well as consider opportunities in other regions of the US. The goal of the system would be to serve as the platform that matched workers, industry, government, non-profits, and educational institutions to the skills currently in demand.

³ <https://www.dol.gov/general/topic/training/adulttraining>

Data Exploration and Preparation

Data Processing Pipeline

The foundation of the proposed system is the data processing pipeline. The pipeline, diagramed below, ingests the job listing data files, engineer features utilized by the machine learning models downstream, and applies those models, warehousing their predictions and extractions to expose them via business intelligence dashboards for analysis.

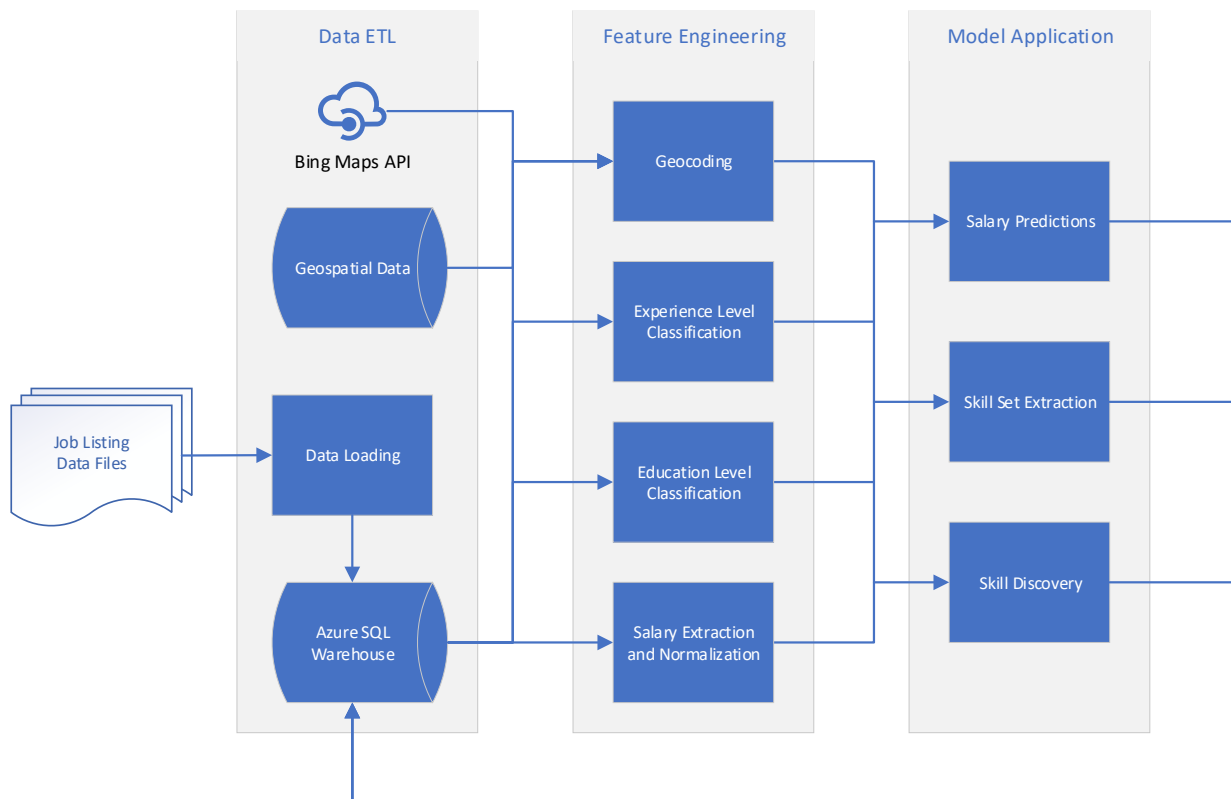


Figure 2: Skill Match Data Processing Pipeline

Data Extraction, Transformation, and Loading (ETL)

Job listing data files are multi-line, comma-delimited text files (CSV) with double quote field qualifiers.

Job listings are broken out by job role and geographic region per file, with the file name also containing a date stamp indicating the date of collection as illustrated here:

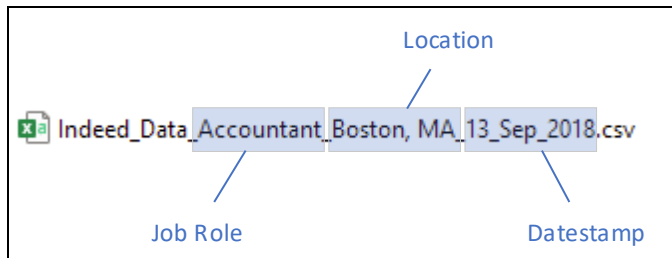


Figure 3: Job Listing Data File Names

The load script responsible for parsing job listings from the CSV files, also parses individual component values from the file names and associates them with the postings when loading them into the data model. These values are used to populate the **JobRole** and **JobPosting** tables.

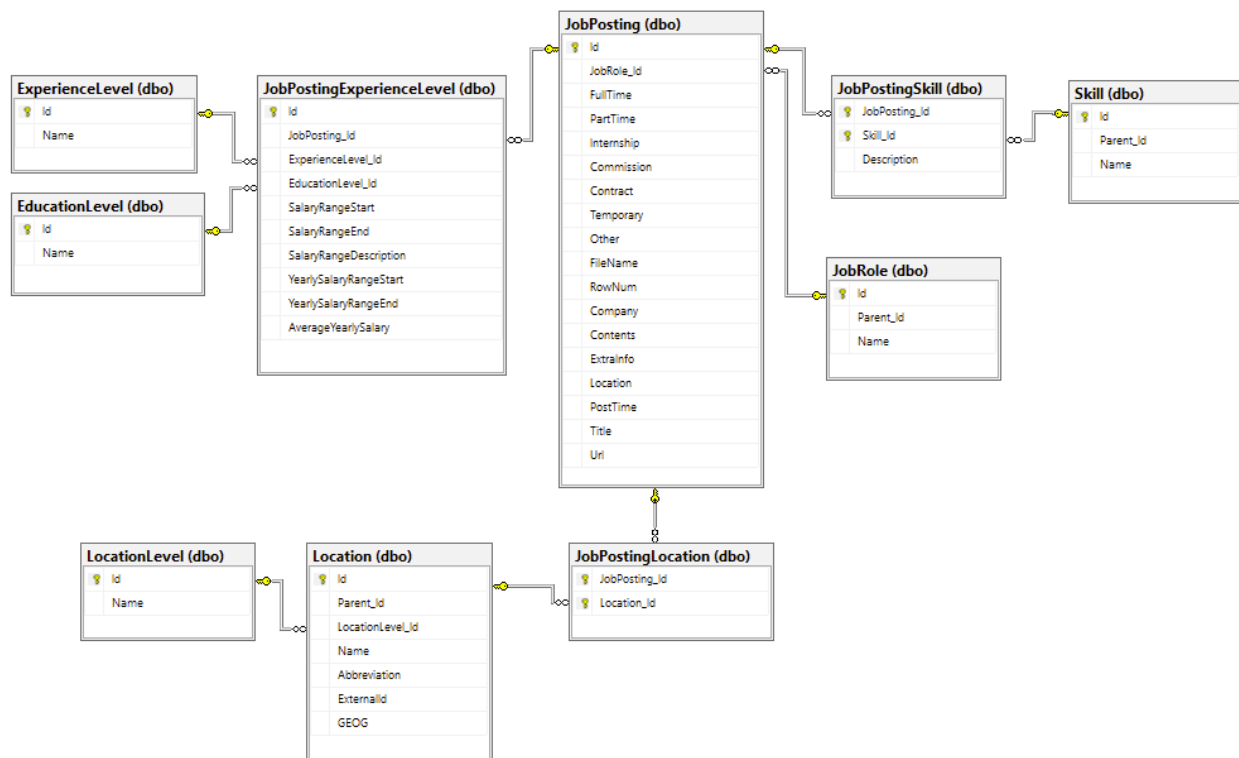


Figure 4: Skill Match Data Model

Feature Engineering

Geocoding

Central to the proposed business objective of the Skill Match system, is the geocoding of job listings, allowing for geospatial analysis of the model predictions and skill extractions. The **JobPosting.Location** field values for the postings are semi-structured, offering a minimum of state and location names, and sometimes also providing a zip code. In some instances, the location values are not the names of cities, towns, or counties, but are informal location names such as “Research Triangle Park, NC”.

To enable the geocoding and geospatial analysis of job postings, a geospatial data store was created within the data warehouse. The **LocationLevel** and **Location** tables store a hierarchical representation of every state, county, city, town, census designated place (CDP), and other administrative entities in the US. The Bing Maps REST API is employed to translate the informal location names, neighborhoods, and points of interest to administrative entities. The geocoding API allows for querying via names, such as those contained in the job listings data set and returns the centroid for a given area, along with the top-level administrative district, second-level administrative district, locality, and the type of entity.

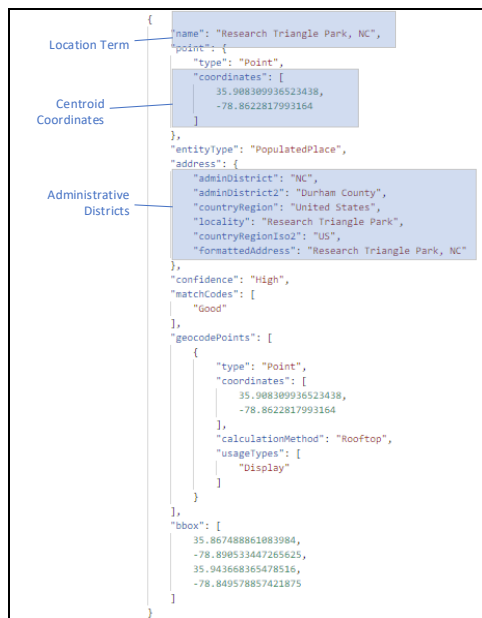


Figure 5: Bing Maps Geocoding API Sample Results

Upon full geocoding of a job posting, the **JobPostingLocation** table is populated.

Experience Level & Education Level Classification

Two key features analyzed in support of the salary prediction model are the experience and education level classifications. These fields are not provided with the job postings and thus were created and derived from the `JobPosting.Title` and `JobPosting.Contents` fields. Upon initial insertion into the system, job postings were given default values of “Unknown” for both features. A script was executed with complex regular expressions to identify commonly used terms for both fields.

For experience levels, job postings were classified into one of three levels: Junior, Mid, and Senior. Job posting titles were first analyzed for titles commonly denoting senior positions, then for junior positions if senior titles were not found, and finally for mid-level titles.

Senior Titles

```
(?<SeniorTitles>\bSenior\b|\bSnr\b|\bSr\b|\bManager\b|\bMngr\b|\bMgr\b|\bSupervisor\b|\bExecutive\b|\bExec\b|\bLead\b|\bDirector\b|\bIV\b|\bV\b|\b[4-6]\b)
```

Junior Titles

```
(?<JuniorTitles>\bJunior\b|\bJr\b|\bEntry\b\s*\bLevel\b|\b1\b|\bI(?:!\.T\.)\b)
```

Mid Titles

```
(?<MidTitles>\bStaff\b|\bMid\b\s-*\bLevel\b|\bAssociate\b|\bII\b|\bIII\b|\b[2-3]\b)
```

Figure 6: Experience Level Regex

Like the job level feature creation, a collection of regular expressions was run against the contents of the job postings to extract the educational requirements. A subset of those expressions are shown below.

Associate Degree

```
(?<AssociateDegree>(\bassociate(?:' '*s*s[' '])*|\bAssoc\b){1}\s*(\bor\b|\bof\b|level|Degree|required|preferred){1})
```

Bachelor's Degree

```
(?<BachelorsDegree>(\bBachelor(?:' '*s*s[' '])*|Bachelor|\bBA\b|\bBS\b){1}\s*(Degree|required|preferred)*)
```

Master's Degree

```
(?<MastersDegree>(\bmaster(?:' '*s*s[' '])*|\bMS\b|\bMA\b){1}\s*(\bor\b|\bof\b|level|Degree|required|preferred){1})
```

Figure 7: Education Level Regex Sample

Salary Extraction and Normalization

The final feature required in support of salary predictions are the salary ranges themselves. These ranges, along with the experience and education levels comprise the data populated in the `JobPostingExperienceLevel` table.

The salary ranges were included in the source data's `JobPosting.ExtraInfo` and `JobPosting.Contents` fields. The salaries were expressed in a number of ways, sometimes only offering a single figure, other times a range complete with upper and lower bounds. There was also variance in the units salaries were expressed in, with hourly, weekly, monthly, and annual figures. This required extracting the salary units and normalizing the salary ranges into standard annual figures for use in the Skill Match salary prediction model. A standard annual work schedule of 40 hours per week for 50 weeks per year was assumed. Therefore, hourly rates extracted from the job descriptions were multiplied by 2000 to arrive at an estimated annual salary. Weekly rates were multiplied by 50 and monthly salaries were multiplied by 12.

```
\$\s*(?<startingSalary>\d{0,3}\.?\d{0,2}\s*k\b|(\d{1,3},(\d{3},)*\d{3}|\d+)(\.\d{2})?)+((?<delimiter>(\s*(-|to|and|a)\s*)*)\$\s*(?<endingSalary>\d{0,3}\.?\d{0,2}k|(\d{1,3},(\d{3},)*\d{3}|\d+)(\.\d{2})?))*(\s*(a|an|per|\\|/)*\s*(?<unit>hourly|hour|hora|hr.|hr|day|d\b|weekly|week|wk|monthly|month|mo\b|yearly|year|y\b|annually|annual))*(?![\$\d.+-]*\s*(m+\b|mill|b\b|bill|t|T))
```

Figure 8: Salary Extraction Regex

```
if (Regex.Match(Unit, @"(hourly|hour|hora|hr.|hr)", RegexOptions.IgnoreCase | RegexOptions.CultureInvariant).Success && currency < 2000)
    return currency.Value * 2000;

if (Regex.Match(Unit, @"(day|d\b)", RegexOptions.IgnoreCase | RegexOptions.CultureInvariant).Success)
    return currency.Value * 250;

if (Regex.Match(Unit, @"(weekly|week|wk)", RegexOptions.IgnoreCase | RegexOptions.CultureInvariant).Success)
    return currency.Value * 50;

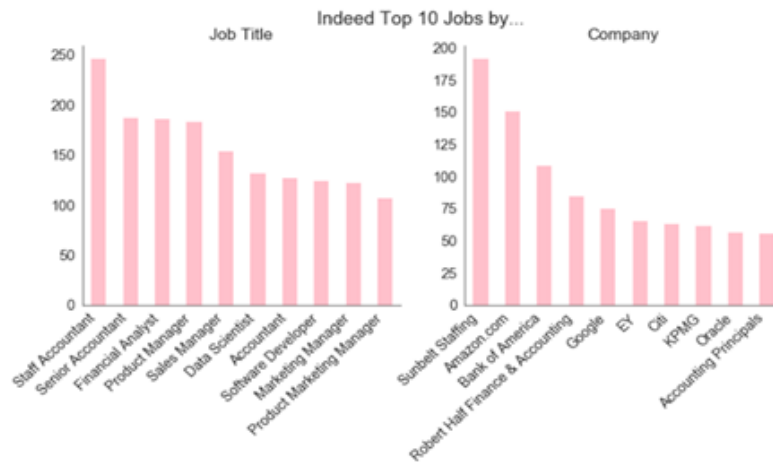
if (Regex.Match(Unit, @"(monthly|month|mo\b)", RegexOptions.IgnoreCase | RegexOptions.CultureInvariant).Success)
    return currency.Value * 12;

if (Regex.Match(Unit, @"(yearly|year|y\b|annually|annual)", RegexOptions.IgnoreCase | RegexOptions.CultureInvariant).Success)
    return currency.Value;
```

Figure 9: Salary Normalization Code (C#)

Exploratory Data Analysis

In order to gain insights into the patterns of the data, Exploratory Data Analysis was conducted on the job postings to analyze the top 10 job titles that were in demand as well as the top 10 companies that were hiring for positions around the United States.



The largest count of job postings observed was for Staff Accountants, followed by Senior Accountants and Financial Analysts. While the top hiring companies include Sunbelt Staffing, Amazon, and Bank of America.



Using plotly, a Choropleth map of the United States was constructed in order to analyze job postings by State. California, Texas and New York ranked amongst the top 3 States with over 2,000 jobs being posted while 24 of the other States accounted for job postings ranging between 10 to 2000

Removal of Punctuations and Conversion to Lower-Case

To prevent two words that are the same but do not have the same capitalization being recognized as two different tokens, all words were set to lower-case for most analysis tasks. This was not performed for tasks where part-of-speech tagging was required.

Tokenization

Tokenization was done prior to lemmatization and stop-word removal using NLTK's word tokenize function.

Lemmatization

Using lemmatization, we modified tokens with similar meaning and changed them all to the lemma, the base or dictionary form of the word. NLTK's WordNet lemmatizer was used in most cases.

Test Pre-Processing on an Accounting Job Post

After the above tasks were done, we tested the processed data on a single row, which had the title 'Accounting'.

```
import wordcloud
from wordcloud import WordCloud, ImageColorGenerator
import matplotlib.pyplot as plt
import numpy as np
from PIL import Image

wc = WordCloud(background_color="white").generate(text_clean)
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```



Text Clustering: Quick Insights from Unstructured Data

TF-IDF Based Classification TD-IDF, also known as Term Frequency-Inverse Document Frequency is a statistical method to reflect the importance of a word to a document in a collection.

We have reused the Pre-Processing methods in the Pre-processing section. As performed previously, we have generated the TF-IDF and reduced the dimension through Singular Value Decomposition (SVD) to 1500 columns.

```
explained_variance = svd.explained_variance_ratio_.sum()
print("Explained variance of the SVD step: {}".format(int(explained_variance * 100)))
```

Explained variance of the SVD step: 68%

K-Means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

```
: from sklearn.cluster import KMeans
  from sklearn import metrics
  num_clusters = 7

  for num in [num_clusters]:
      km3 = KMeans(n_clusters=num, init='k-means++', max_iter=1000, n_init=1, random_state=1)
      %time km3.fit(X_lsa)
      # The higher the better (-1 to 1)
      print("Clusters: {}".format(num))
      print("Silhouette Coefficient for clusters: %.3f"
            % metrics.silhouette_score(X_lsa, km3.labels_))
```

Wall time: 8.23 s

Clusters: 7

Silhouette Coefficient for clusters: 0.028

```
def print_terms(cm, num):
    original_space_centroids = svd.inverse_transform(cm.cluster_centers_)
    order_centroids = original_space_centroids.argsort()[::-1]
    terms = vec_tfidf.get_feature_names()
    for i in range(num):
        print("Cluster %d:" % i, end='')
        for ind in order_centroids[i, :10]:
            print(' %s' % terms[ind], end='')
        print()

print_terms(km3, num_clusters)
```

Cluster 0: accounting financial accountant tax reporting account finance monthly reconciliation analysis
Cluster 1: patient care physician health school medical clinical social hospital family
Cluster 2: data business client management financial ability risk service security project
Cluster 3: product marketing customer market business strategy management development sale manager
Cluster 4: sale marketing customer business manager product account campaign ability company
Cluster 5: software web development application developer database design technology system sql
Cluster 6: project construction design engineering mechanical engineer civil management client required

Salary Predictions

In order to explore the job postings data thoroughly, making accurate salary predictions that are based on known salaries can be useful to companies and job seekers. Job seekers will be able to match their skills, years of experience and location along with their education background to secure a job. This model will serve as a guide for offering competitive compensation to existing and future employees while controlling payroll expenses.

The main features used for Salary prediction were:

1. Years of Experience
2. Experience Level
3. Education Level
4. Industry
5. Job City
6. Job State

The Accountant job role had the highest counts of salary ranges extracted from the data set, thus these salaries were used in the training and testing of the prediction model. Given the likelihood of salaries to vary greatly by experience level, the first step in the model was to bucket salaries based upon the four different levels (unknown, junior, mid, and senior).

Similarly, salary ranges were bucketed by education level, separating them into Masters, Bachelors, Associate, High School and unknown.

To create a basic training model, 20% of the training data was split into testing data so that the model could be tested with data for which the salaries are already known. The training data was used to fit a simple linear regression model. After the baseline model was created, we were then able to predict salaries. The first five predicted values were \$88,724, \$66,148, \$68,536, and \$76,380.

Three additional models were run: linear regression with second order polynomial transformation, random forest regression, followed by ridge regression. The most common way to evaluate the overall fit of a linear model is by R-Squared value, but the R-Squared value does not always measure the usefulness of the model. Thus, selecting the model with the highest R-squared is not the most reliable approach for choosing the best linear model. Instead, the mean squared error was used in the evaluation.

<i>Model</i>	<i>R-Squared</i>	<i>Mean Squared Error</i>
Linear Regression	.75	95.65
Linear Regression w/ 2 nd order polynomial transformation	.89	39.61
Random Forest	.54	218.69
Ridge Regression	.88	46.41

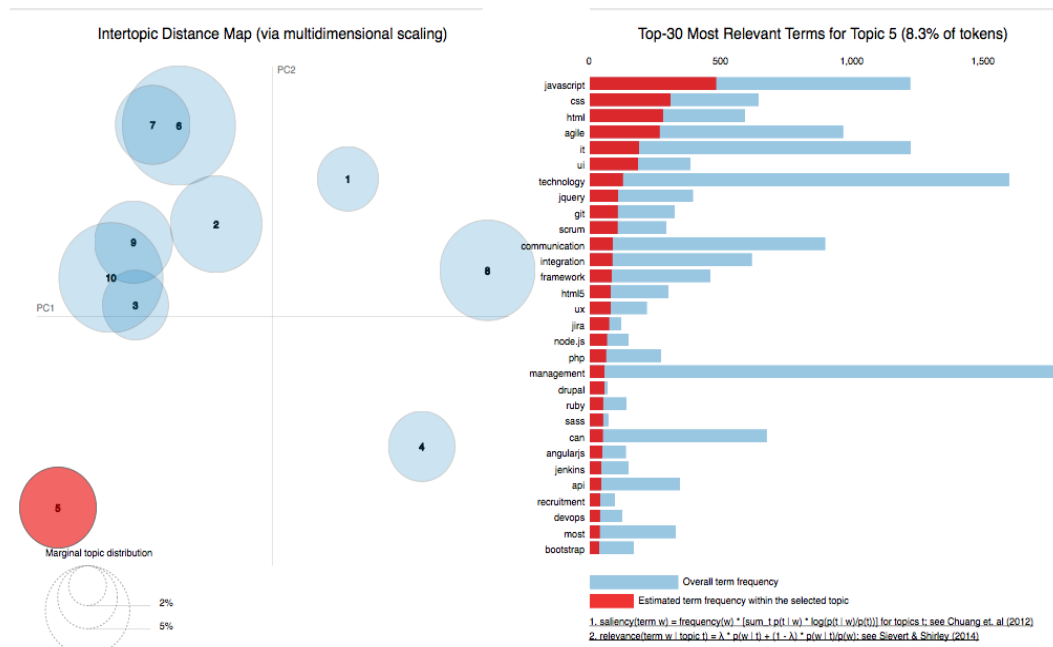
A larger mean squared error indicates data values are dispersed widely around its central moment. A smaller mean squared error indicates otherwise and it is definitely the preferred and/or desired choice as it shows that our data values are dispersed closely to its central moment, which is usually ideal.

Applying second order polynomial transformation to the features gave the most accurate predictions with the least error when using a linear regression model. The result was a mean squared error of 39.61 with 89% accuracy rate. This model can be used as a guide when determining salaries since it shows reasonable predictions when given information on years of experience, job type, job posting city, job posting state and education level.

Skill Extraction/Discovery

Topic Modeling and Extraction

Latent Dirichlet Allocation (LDA) was used to classify the text data in the “Content” column of the dataset to a particular topic. It builds a topic per job description model and words per topic model, modeled as Dirichlet distributions. Each job description is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words. The model assumes that every chunk of text fed into it will contain words that are somehow related. The model also creates a bag of words that contains the frequency of the word in relation to a specific topic.



The output from the model topics are categorized by a series of words. The LDA model, in general, doesn't give a topic name to the words and is meant to be interpreted manually.

```
'0.205*software" + 0.039*computer" + 0.028*python" + 0.025*security" + 0.025*c++" + 0.024*programming" + 0.022*c" + 0.021*java" + 0.020*technology" + 0.020*linux"
```

The output generated from Topic 5 is associated to the job-posting of a Software Engineer.

In order to further refine the skills associated with each job posting a pre-assigned skills data set with close to 8000 entries was used as the training data in order to compare with a random job description from the dataset.

```
' Managing master data, including creation, updates, and deletion. Managing users and user roles. Provide quality assurance of imported data, working with quality assurance analyst if necessary. Commissioning and decommissioning of data sets. Processing confidential data and information according to guidelines. Helping develop reports and analysis. Managing and designing the reporting environment, including data sources, security, and metadata. Supporting the data warehouse in identifying and revising reporting requirements. Supporting initiatives for data integrity and normalization. Assessing tests and implementing new or upgraded software and assisting with strategic decisions on new systems. Generating reports from single or multiple systems. Troubleshooting the reporting database environment and reports. Evaluating changes and updates to source production systems. Training end users on new reports and dashboards. Providing technical expertise on data storage structures, data mining, and data cleansing.'
```

When testing the model, these were the top skills that were extracted after matching with the pre-defined skills dataset.

```
[ 'analysis',
  'reporting',
  'security',
  'warehouse',
  'reporting',
  'normalization',
  'software',
  'troubleshooting',
  'reporting',
  'database',
  'training' ]
```

In order to expand on the extracted words in order to get some context to the skills, Bootstrapping was done. A relatively new package, Skills -ML comes with a bootstrapping extraction technique whereby the skills as well as the sentences associated with those skills are shortlisted.

{1: 'About Onnit:\n\nOnnit is an Austin, TX-based health and wellness brand focused on encouraging a peak level of human performance through the best in nutritional supplementation, health-conscious foods, and unconventional fitness equipment and training. We are rapidly growing through support from professional athletes, medical practitioners, and our thousands of customers. We are proud of our Austin roots and have established a world-class gym next door to our headquarters, as well as acquired local favorite yoga studio Black Swan Yoga. We're excited to be named a Best Place to Work in Austin by Austin Business Journal, and we'd love to show you why.\n\nAbout the Position:\n\nOnnit Labs is seeking an experienced, detail-oriented Staff Accountant to join our team. In this position, the Staff Accountant's primary responsibility will be processing payroll on a bi-weekly frequency for multiple companies as well as recording all payroll related transactions into the general ledger. This position will report directly to the Controller and will be responsible for assisting in month-end close, reconciliations, analysis of accounts, journal entries and various special projects. In this fast-paced environment, a high level of organization and an ability to multitask projects is crucial for success. This position is integral to the success of the accounting department and the company as a whole.\n\nResponsibilities:\n\nPrepare and process payroll on a bi-weekly basis for parent company and its subsidiaries.\n\nResponsible for accurate posting of all payroll activity to the General Ledger.\n\nSupport the accounting and administration activities associated with all employee benefits - health, vision, dental, 401k, and other company-specific benefits.\n\nAssist in month-end closing duties including accruals, identifying discrepancies, proposing adjusting journal entries, as well as other general closing support.\n\nAnalyze current policies and procedures in order to develop and implement changes for improved efficiency.\n\nEnsure compliance with internal controls.\n\nAssist with external and internal audits as needed.\n\nFacilitate miscellaneous payroll reporting for all external and internal users, as requested.\n\nTroubleshoot and investigate discrepancies, including but not limited to payroll.\n\nCoordinate the distribution of W-2s.\n\nPerforms all duties and responsibilities in a timely and effective manner to ensure all deadlines are met.\n\nPerform special projects as assigned.\n\nRequirements\n\nQualifications:\n\nBachelor's degree from a four-year college or university with an emphasis in Accounting/Finance or related field required.\n\nBackground check required prior to hire\n\nExperience:\n\nMinimum 2-3 years of Accounting/Finance experience required.\n\nExperience using an ERP system such as NetSuite will be a plus.\n\nBenefits\n\nFull medical, dental, and vision benefits\n\nBasic Life Insurance\n\n401(k) eligibility with company matching\n\nFlexible Vacation and time off policy\n\nPaid holidays\n\nCompetitive compensation\n\nNutritional reimbursement\n\nFringe benefits including free access to Onnit Academy gym and Black Swan Yoga studios; weekly co-pay massages; co-pay healthy meals; among many others'}

The previous job description is narrowed down to the skills associated with the job:

skill name: Prepare and process payroll on a bi-weekly basis for parent company and its subsidiaries.	
skill name: Responsible for accurate posting of all payroll activity to the General Ledger.	
skill name: Support the accounting and administration activities associated with all employee benefits – health, vision, dental, 401k, and other company-specific benefits.	
skill name: Assist in month-end closing duties including accruals, identifying discrepancies, proposing adjusting journal entries, as well as other general closing support.	
skill name: Analyze current policies and procedures in order to develop and implement changes for improved efficiency.	
skill name: Ensure compliance with internal controls.	
skill name: Assist with external and internal audits as needed.	
skill name: Facilitate miscellaneous payroll reporting for all external and internal users, as requested.	
skill name: Troubleshoot and investigate discrepancies, including but not limited to payroll.	
skill name: Coordinate the distribution of W-2s.	
skill name: Performs all duties and responsibilities in a timely and effective manner to ensure all deadlines are met.	
skill name: Perform special projects as assigned.	
skill name: Bachelor's degree from a four-year college or university with an emphasis in Accounting/Finance or related field required.	
skill name: Background check required prior to hire	

Recommendation/Conclusion

Shown below are three separate views from a Tableau Workbook visualizing the output generated from the Skill Match system. The analysis afforded by this system allows governmental and non-governmental organizations provide better education pathways for the citizens to gain skills in demand. Likewise, for current and future job seekers, it is recommended that if they have interest, they should not hesitate to equip themselves with skillset and qualifications shown by the system.

