

## CASE STUDY: JOB SALARY PREDICTOR

We will do a simple case study on using text analytics to assist job seekers negotiate salaries by predicting what salary to expect for a certain job description at a given location.

### Opening prompt / problem statement

There are tens of thousands of jobs on online job boards that invite job seekers to apply for jobs without providing an estimate of the compensation they provide – this is further validated when scraping for salaries online. This poses a lot of problems for job seekers, especially new graduates, applying for a job role or when trying to negotiate the compensation.

### Context and opportunity

One approach is to train machine learning models to predict salaries for a job posting at a certain location in the United States. The model has to use the text descriptions used when advertising the job, but also ‘discrete’ features such as location and title. Job seekers face this problem every day in their job hunt which provides a strong motivation to build an efficient and reliable model that finds the right compensation for the right job and level of experience.

### Technical Details

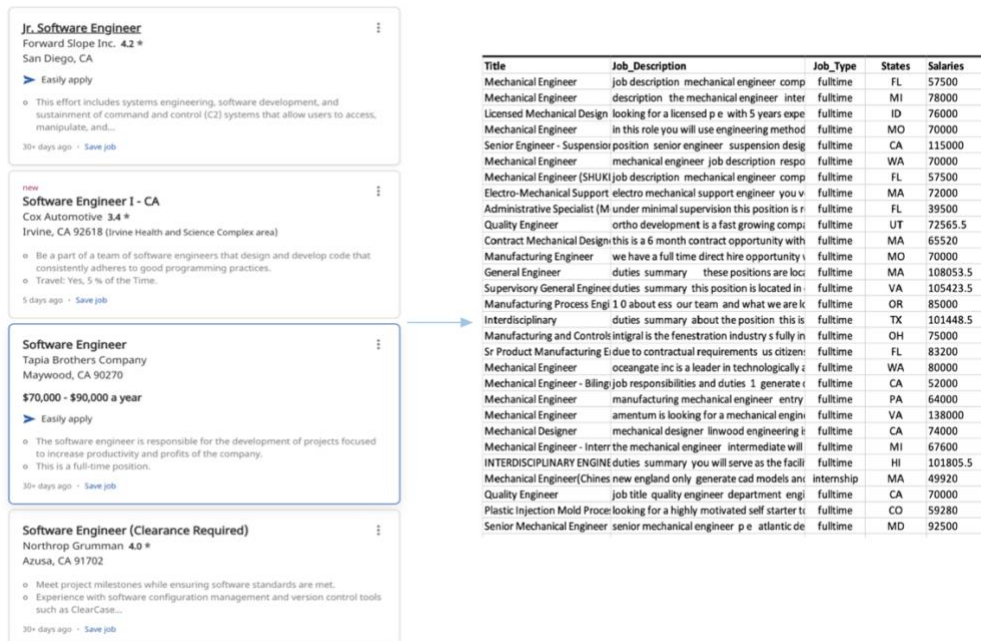
In order to predict the salary from the set of predictors, we need to collect a lot of data to train the machine learning models. Furthermore, we want to work with raw data from the real-world; hence one option would be to scrape data off Indeed.com or other popular job boards with a variety of jobs. One could use the Selenium web scraper with search strings ranging from ‘software engineering’ to ‘data scientist’ to scrape (potentially) tens of thousands of postings.

Even after scraping, data cleaning is an important challenge that needs to be dealt with (see case question ii). Examples of data cleaning steps include removing punctuation, new line characters, extra spaces/tabs and stopwords; converting all formats to salary (in \$) per year

## CASE STUDY: JOB SALARY PREDICTOR

using RegEx and conversion rules; standardized ‘closed world’ fields such as states and countries, and attempt to correct spelling errors.

Suppose that the final data, after scraping, cleaning and preprocessing before vectorization looks like the following:



Title	Job_Description	Job_Type	States	Salaries
Mechanical Engineer	job description mechanical engineer comp	fulltime	FL	57500
Mechanical Engineer	description the mechanical engineer inter	fulltime	MI	78000
Licensed Mechanical Design	looking for a licensed p e with 5 years expe	fulltime	ID	76000
Mechanical Engineer	in this role you will use engineering method	fulltime	MO	70000
Senior Engineer - Suspension	position senior engineer suspension desg	fulltime	CA	115000
Mechanical Engineer	mechanical engineer job description respo	fulltime	WA	70000
Mechanical Engineer (SHUKI)	job description mechanical engineer comp	fulltime	FL	57500
Electro-Mechanical Support	electro mechanical support engineer you v	fulltime	MA	72000
Administrative Specialist (M	under minimal supervision this position is r	fulltime	FL	39500
Quality Engineer	ortho development is a fast growing compi	fulltime	UT	72565.5
Contract Mechanical Design	this is a 6 month contract opportunity with	fulltime	MA	65520
Manufacturing Engineer	we have a full time direct hire opportunity i	fulltime	MO	70000
General Engineer	duties summary these positions are loci	fulltime	MA	108053.5
Supervisory General Engine	duties summary this position is located in	fulltime	VA	105423.5
Manufacturing Process Engi	1 0 about ess our team and what we are k	fulltime	OR	85000
Interdisciplinary	duties summary about the position this is	fulltime	TX	101448.5
Manufacturing and Controls	Integral is the fenestration industry s fully in	fulltime	OH	75000
Sr Product Manufacturing E	due to contractual requirements us citizen:	fulltime	FL	83200
Mechanical Engineer	oceangate inc is a leader in technologically i	fulltime	WA	80000
Mechanical Engineer - Biling	job responsibilities and duties 1 generate i	fulltime	CA	52000
Mechanical Engineer	manufacturing mechanical engineer entry	fulltime	PA	64000
Mechanical Engineer	amentum is looking for a mechanical engin	fulltime	VA	138000
Mechanical Designer	mechanical designer linwood engineering i	fulltime	CA	74000
Mechanical Engineer - Interr	the mechanical engineer intermediate will	fulltime	MI	67600
INTERDISCIPLINARY ENGINE	duties summary you will serve as the facili	fulltime	HI	101805.5
Mechanical Engineer(Chines	new england only generate cad models an	internship	MA	49920
Quality Engineer	job title quality engineer department engi	fulltime	CA	70000
Plastic Injection Mold Proce	looking for a highly motivated self starter t	fulltime	CO	59280
Senior Mechanical Engineer	senior mechanical engineer p e atlantic de	fulltime	MD	92500

Many job titles and descriptions, despite being different, are actually very similar in practice (e.g., ‘Jr. Software Engineer’ vs. ‘Software Engineer I’). One way to ‘cluster’ such similar jobs is by using clustering. Recall the k-means algorithm we studied in lecture. One of the issues with k-means is that we have to specify k in advance. Also, it is not exploratory.

An alternate approach is to use topic modeling, with different values for k (the number of topics). An advantage of this approach is that it shows us the word clouds associated with each topic:

## CASE STUDY: JOB SALARY PREDICTOR

Topic #2: (Software Engineering)

software, database, agile, technology, sql, cloud, software development, etl, integration, computer, intelligence, architecture, infrastructure, computer science, programming, business intelligence, big data, data engineering, scrum, warehouse, management, data warehouse, automation, business requirements, security, data management, data processing, implementation, linux, innovation

Topic #3: (Mechanical Engineering)

manufacturing, product development, mechanical engineering, solidworks, communication, assembly, cross-functional, root, cad, gmp, analytical, technology, automotive, manufacturing engineering, written communication, troubleshooting, problem solving, chemical engineering, lean manufacturing, process improvement, curriculum, interpersonal, solid, mathematical, automation, plc, heat, technical support, inspection, supervision

One advantage of the topic model is that we can use the topic to which a job has been assigned as a type of feature. This ensures that jobs that are ‘topically’ similar have a feature in common. We could also use word embeddings and other features to ‘vectorize’ the job description or even the titles.

Once we have processed the text, we need to deal with the discrete features, perhaps by normalizing them. Next, we need to train a machine learning model with some training data (using known salaries for jobs, either proprietarily, or by using select data from sources like Glassdoor). We then need to evaluate it on all our scraped data, or potentially allow a user to get predictions by posting job details on a Web application user interface:

# CASE STUDY: JOB SALARY PREDICTOR

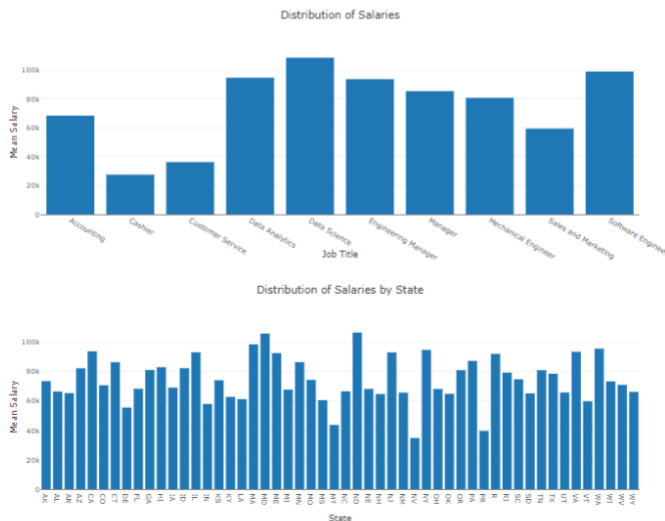
[Job Salary Predictor](#) [Final Project - ISE 540 Test Analytics](#) [Contact](#)

## Job Salary Predictor

Analyzing job descriptions for predicting salary

Predict Salary

## Overview of Training Dataset



## Case Questions

- In *Technical Details*, we mentioned search strings ranging from 'software engineering' to 'data scientist'. Give a few more examples of what the search strings could be, and whether you would want to limit by technical fields.
- Suppose you scraped a CSV file that contains data along the lines provided below. How would you clean this dataset?

## CASE STUDY: JOB SALARY PREDICTOR

46–49 an hour	Goleta, CA 93117
9, 588–13,460 a month	Fresno, CA 93721 (Central area)
\$8,000 a month	Industry, CA 91789
From \$17 an hour	None
90, 000–95,000 a year	Los Angeles, CA 90028 (Hollywood area)
32–50 an hour	Boston, MA
135, 000–185,000 a year	None
75, 000–100,000 a year	None

iii) We suggested clustering as one way to deal with jobs that look ‘different’ but are actually very similar. What are some metrics that can be used to evaluate clustering? Instead of using topic model or even k-means, would it make sense to use hierarchical or agglomerative clustering? What are the pros and cons of such an approach?

iv) Describe in detail how you might use an ensemble machine learning model to learn something that works across many sectors and industries.

v) Why would you want to use word embeddings vs. an approach like tf-idf in a problem such as this one? Hint: think about whether you might encounter words that are not present in the corpus you obtained to get your tf-idf model. Why is there particularly good reason to suspect this problem might arise in this domain?

vi) Suppose you have a model that works quite well right now. How might you extend it to work over longer time periods (say, 2-5 years) without re-training?

vii) Comment on how you would extend the model or pipeline to deal with non-US geographies, and non-English job postings.

viii) Comment on the metrics for success on the various stages of the pipeline presented in *Technical Details*.