Research Proposal

Team 28

Research Motivation

Among the many things that haunt a graduate student's mind, finding a job is probably the most important and monstrous. While an ideal job is the intersection of three sets – what one loves, what one is good at, and what the society values – an answer to the third question is enough to guarantee a good pay. So here comes our mundane but vital research question: **what features will guarantee you a good pay in the job market?**

The next thing we want to consider is how we are going to answer this question. Consider ourselves data scientists, we would say: use data! Ideally, if we can extract features from job requirements and find correlations between these features and the salaries offered, we can have a satisfactory answer to this question.

Data and Definitions

In order to precisely predict the salary of a certain job from its characteristics, we need the data of the salary, job industry, titles and other factors, so we searched online and found it on NYC Open Data.

Our dataset contains 1658 rows and 26 variables; each row looks like below:

Each row represents a specific job post. It contains Job.id, agency, posting type, number of vacancies to be fill, business title, civil service title, title code number, level, salary range (from the lowest salary to the highest salary), salary frequency, work location, division/work unit, job description, minimum quality required, preferred skills, additional information, how to apply, hours/shift, work location, recruitment contact, residence requirement, posting date, post until, posting updated date and process date.

Among them salary range from and salary range to are numeric variables showing the lowest and the highest salary for each posting job; job description, minimum quality requirements are description of a specific position which we expect to extract characteristics of the jobs from.

Research Questions Refined

Based on the data set we find, we formalize our research questions as below.

- What features are strong predictors of the salary? There are basically two feature sets: one contains those protowords derived from the linguistic context of job descriptions, the other contains non-linguistic covariates in the data set.

Research Design

In this part, we will elaborate on what our dependent and independent variables are, how they will be constructed, and what models are good candidates to assess the relationship between them.

- Dependent Variable: salary level
- Independent Variables: linguistic features and non-linguistic features

❖ Variable Construction

For the dependent variable *Y*=salary level, we will calculate its mean and bin it into *n* classes where each matches to a set of jobs.

For each observation, we will combine the texts from “job description”, “preferred skills” and “minimum requirements”, tokenize each sentence in the newly formed text, remove punctuations and common stop words, exclude words generally used in job descriptions, compute unigrams, bigram, and trigrams, match them to phrases from selected dictionaries, stem the words, and build up a word bag for each job.

We will merge all the word bags and compute a corpora of words. Each word *w* used at least once in the text context by a job is assigned a score for each of the classes based on the conditional probability of the class given the word. For each class, we select *k* words of highest scores. The *n*k* words collected from all ranges are used as linguistic features for a job.

For each word variable a job is assigned the score computed by:

$$\frac{number\ of\ times\ this\ word\ used\ by\ the\ job}{addition\ of\ number\ of\ times\ all\ words\ used\ by\ the\ job}$$

In addition to the linguistic features, we are going to compute non-linguistic features such as post length, reading score, and location features etc. We will build up a metric to compare performances using three features sets including *L*=linguistic features, *F*=non-linguistic features and *L* U *F*=all features.

❖ Model Selection

First we will estimate a linear regression of salary midpoints on each set of predictors. Therefore this model is parameterized succinctly by the number of salary brackets we use to generate the protowords in the PU model. We will then estimate three categorical response models: an ordinal generalized linear model (a discriminative model), a naive Bayes classifier (a generative model), and a decision tree. For parsimony and simplicity of interpretation, we will use the same salary categories as in the corresponding PU model. Therefore the categorical models are more dependent than the continuous model on the number of categories we choose. Moreover, quantizing or binning the response without good reason amounts to throwing away information, but it is not that clear the full range of variation in salary contains **useful** information. In this case, binning the response variable could produce a better model by focusing on broader patterns of association, rather than fitting a model to variation that we know we probably cannot explain with our data.

Each model takes a different conceptual approach. Linear regression, even when its classical foundations fail, always produces a first-order approximation of association between the response and the predictors. Among the categorical response models, the ordinal GLM is the closest analogue to linear regression, in that the ordering of the response is preserved and the model parameters can be cleanly interpreted. The naive Bayes classifier abandons ordinality, which would make sense if the predictors of high-wage jobs are not ordinally "larger" than the predictors of low-wage jobs, as our PU setup allows for and in some sense imposes. Finally, the decision tree abandons linearity in the predictors (naive Bayes is linear in log-probability space): there is no obvious reason why the model should be linear and the decision tree is conceptually our "most general."

Our goal is to determine which characteristics of a job posting, if any, are meaningful predictors of wage. High predictive accuracy would suggest highly informative predictors, and in this case a hi

Potential Outcomes & Visualizations

❖ Outcomes

- What results are we going to get? How will we present our results?

❖ Visualizations

- Text Cloud: Shows the results of our model, which characteristics are valued for NYC Government jobs and how important they are.
- Salary Map: Shows how salary connected to location of the job. (Plot the job offerings on the map with the color representing salaries.)
- Result application: Shows how future salaries of Columbia students will vary across disciplines. (Based on Columbia Map)----a funny demo of our prediction results (See Appendix for a fake example)

Remaining Questions

We are losing information by using the mean of the salary, is there any way we can use the range, or both bonds of the salary(lower and higher) , because there might be a relationship between range of the salary and the job characteristics, and do not make our model complicated?

Appendix

❖ Data Set

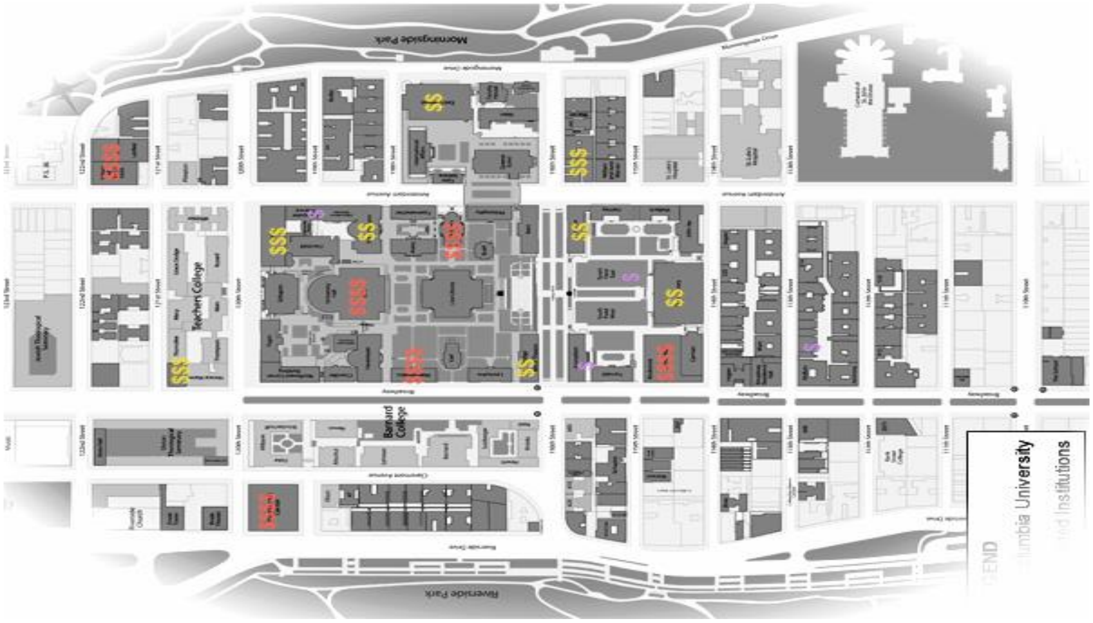
Source https://data.cityofnewyork.us/Business/NYC-Jobs/kpav-sd4t

Table 1: One Row of Our Data

Job ID	139190
Agency	DEPT OF ENVIRONMENT PROTECTION
Posting Type	External
#of Position	1
Business Title	Policy Analyst/Advisor
Civil Service Title	STAFF ANALYST TRAINEE
Title Code No	12749
Level	0
Salary Range From	35538
Salary Range To	49041
Salary Frequency	Annual
Work Location	59-17 Junction Blvd Corona Ny
Division/Work Unit	Executive Management
Job Description	Under the direction of the Treasurer of the New York City Water Board, the selected candidate will work closely with the Treasurer and other executive staff in analyzing revenues and expenses of New York City's Water and Wastewater System, projecting the financial impacts of Water Board policies related to water and wastewater rates and charges, and evaluating the financial aspects of various strategic projects. The candidate will work closely with staff of DEP's Bureaus of the Water Board, Customer Services, Public Affairs and Budget Office. The candidate will also prepare informational presentations for the Water Board and public, promote the Water Board's Rate Hearings, and develop strategies for connecting with New York City's businesses and homeowners who are affected by the Water Board's policies, regulations and rulemaking activities. The candidate will also work with each of DEP's support Bureaus to prepare monthly accountability reports for executive staff.

Minimum Qual Requirements	A baccalaureate degree from an accredited college.
Preferred Skills	Proficiency in Microsoft Excel, Word and PowerPoint.
Additional Information	
To Apply	Click "Apply Now" button
Hours/Shift	
Work Location 1	59-17 Junction Blvd Corona NY
Recruitment Contact	
Residency Requirement	New York City residency is generally required within 90 days of appointment. However, City Employees in certain titles who have worked for the City for 2 continuous years may also be eligible to reside in Nassau, Suffolk, Putnam, Westchester, Rockland, or Orange County. To determine if the residency requirement applies to you, please discuss with the agency representative at the time of interview.
Posting Date	NA
Post Until	NA
Posting Updated	1/2/14 0:00
Process Date	4/22/14 0:00

❖ Visualization – A Fake Example


Morningside Park
Teachers College
Barnard College
Riverside Park
END
mbia University
nd Institutions