CPE 695 APPLIED MACHINE LEARNING

FINAL PROJECT

PERFECT CAREERS - MATCHES YOUR SKILLSETS WITH YOUR JOB PROFILE

Team members: Yuvaraj Ganesh (10415765) and Akanksha Chatra (10418974)

Email: yganesh@stevens.edu and achatra@stevens.edu

Introduction

In this competitive world it is hard to get an interview call, let alone a job. An individual has to apply to a lot of companies before he finally gets an interview call. One of the main reasons for this is that the candidate fails to match the job summary of the company which best fits his resume. In this project we help the candidate to increase his chances of getting an interview call by using various machine learning algorithms. We help the candidate find the best job by scraping data from Indeed.com by scraping over 2000 data every time the program is compiled.

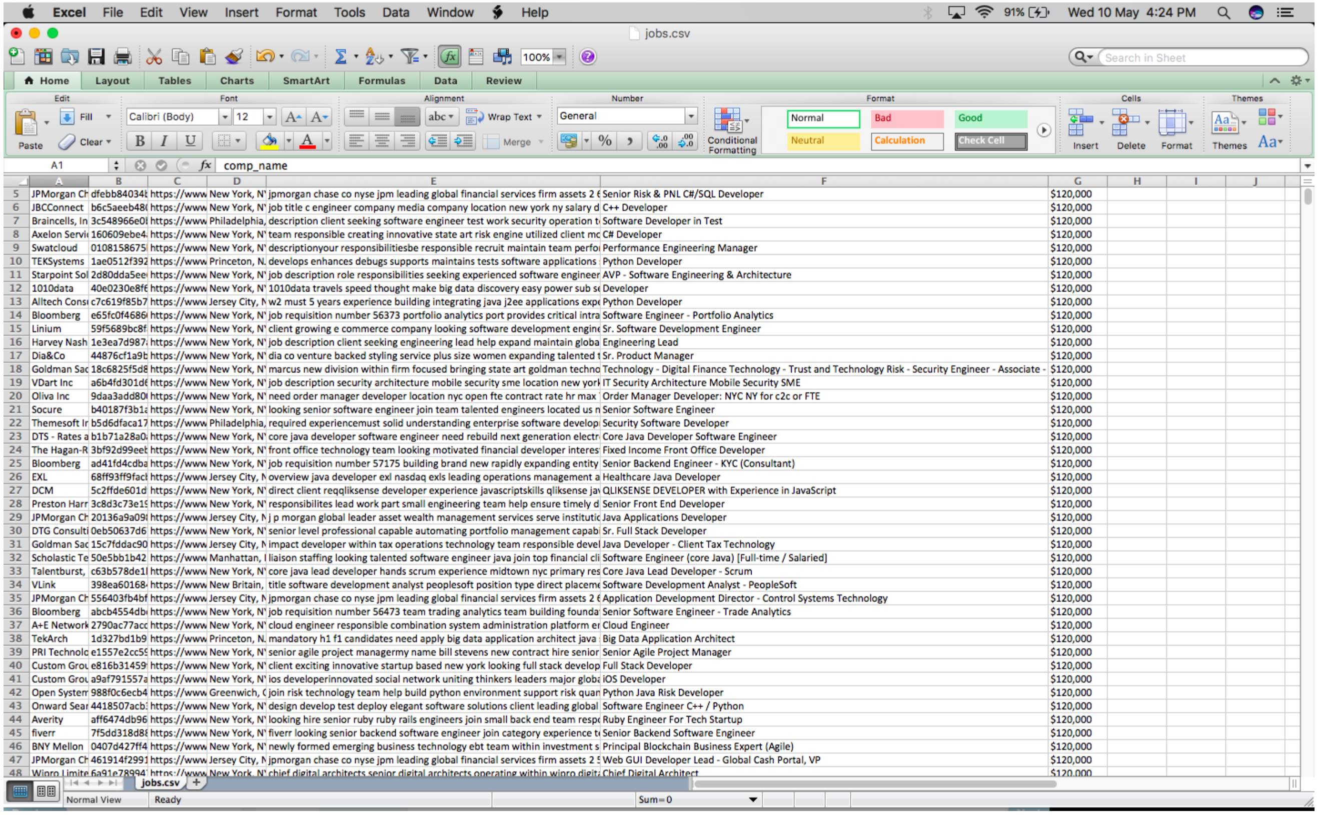
Implementation

In this project, we are implementing the following

- Data Scraping
- Data cleaning
- Predicting the company using machine learning algorithms
- Determining the accuracy

I. Data Scraping

Data scraping mean extracting data from websites. Here we are scraping data from Indeed.com using python. Various parameters like company name, job key, job link, job location, job summary, job title and salary label are scraped from the website. Around 2000 company details is scraped every time the code is compiled. This is achieved by using the BeautifulSoup library. It is an incredible tool for pulling out information from a webpage. We can use it to extract tables, lists, paragraph and we can also put filters to extract information from web pages. BeautifulSoup does not fetch the web page for us. That’s why, urllib is used in combination with the BeautifulSoup library. For the prediction of the model we need an organized dataset which we achieved using pandas libraries. Here, we write the scraped data into a csv file using pandas. Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. The library is a fast and efficient DataFrame object for data manipulation with integrated indexing. It is used to read and write data between in-memory data structures and different formats: CSV and text files, Microsoft Excel, SQL databases, and the fast HDF5 format. Based on the salary range we are scraping data and collecting around 2000 company data. We scrape fresh data every time because using previous data would decrease the efficiency of the model as the company job requirements change on a timely basis or the company job posting has expired. The figure below shows a .csv file containing company features scraped from Indeed.com


Excel File Edit View Insert Format Tools Data Window Help
Wed 10 May 4:24 PM
jobs.csv
Home Layout Tables Charts SmartArt Formulas Data Review
Edit Font Alignment Number Format Cells Themes
Paste Fill Clear Calibri (Body) 12 Wrap Text Merge General
Conditional Formatting
Normal Bad Good Neutral Calculation Check Cell
Insert Delete Format Themes
Search in Sheet
Normal View Ready Sum=0

A1 comp_name

	A	B	C	D	E	F	G
5	JPMorgan Ch	dfebb84034	https://www	New York, N	jpmorgan chase co nyse jpm leading global financial services firm assets 2	Senior Risk & PNL C#/SQL Developer	$120,000
6	JBCConnect	b6c5aeeb48	https://www	New York, N	job title c engineer company media company location new york ny salary d	C++ Developer	$120,000
7	Braincells, In	3c548966e0	https://www	Philadelphia,	description client seeking software engineer test work security operation t	Software Developer in Test	$120,000
8	Axelon Servi	160609ebe4	https://www	New York, N	team responsible creating innovative state art risk engine utilized client mc	C# Developer	$120,000
9	Swatcloud	0108158675	https://www	New York, N	descriptionyour responsibilitiesbe responsible recruit maintain team perfo	Performance Engineering Manager	$120,000
10	TEKSystems	1ae0512f392	https://www	Princeton, N	develops enhances debugs supports maintains tests software applications	Python Developer	$120,000
11	Starpoint Sol	2d80dda5ee	https://www	New York, N	job description role responsibilities seeking experienced software engineer	AVP - Software Engineering & Architecture	$120,000
12	1010data	40e0230e8f6	https://www	New York, N	1010data travels speed thought make big data discovery easy power sub s	Developer	$120,000
13	Alltech Cons	c7c619f85b7	https://www	Jersey City, N	w2 must 5 years experience building integrating java j2ee applications exp	Python Developer	$120,000
14	Bloomberg	e65fc0f4686	https://www	New York, N	job requisition number 56373 portfolio analytics port provides critical intra	Software Engineer - Portfolio Analytics	$120,000
15	Linium	59f5689bc8f	https://www	New York, N	client growing e commerce company looking software development engin	Sr. Software Development Engineer	$120,000
16	Harvey Nash	1e3ea7d987	https://www	New York, N	job description client seeking engineering lead help expand maintain globa	Engineering Lead	$120,000
17	Dia&Co	44876cf1a9b	https://www	New York, N	dia co venture backed styling service plus size women expanding talented t	Sr. Product Manager	$120,000
18	Goldman Sac	18c6825f5d8	https://www	New York, N	marcus new division within firm focused bringing state art goldman techno	Technology - Digital Finance Technology - Trust and Technology Risk - Security Engineer - Associate -	$120,000
19	VDart Inc	a6b4fd301d6	https://www	New York, N	job description security architecture mobile security sme location new york	IT Security Architecture Mobile Security SME	$120,000
20	Oliva Inc	9daa3add80	https://www	New York, N	need order manager developer location nyc open fte contract rate hr max	Order Manager Developer: NYC NY for c2c or FTE	$120,000
21	Socure	b40187f3b1a	https://www	New York, N	looking senior software engineer join team talented engineers located us n	Senior Software Engineer	$120,000
22	Themesoft Ir	b5d6dfaca17	https://www	Philadelphia,	required experiencemust solid understanding enterprise software develop	Security Software Developer	$120,000
23	DTS - Rates a	b1b71a28a0	https://www	New York, N	core java developer software engineer need rebuild next generation electr	Core Java Developer Software Engineer	$120,000
24	The Hagan-R	3bf92d99eel	https://www	New York, N	front office technology team looking motivated financial developer interes	Fixed Income Front Office Developer	$120,000
25	Bloomberg	ad41fd4cdba	https://www	New York, N	job requisition number 57175 building brand new rapidly expanding entity	Senior Backend Engineer - KYC (Consultant)	$120,000
26	EXL	68ff93ff9fac	https://www	Jersey City, N	overview java developer exl nasdaq exls leading operations management a	Healthcare Java Developer	$120,000
27	DCM	5c2ffde601d	https://www	New York, N	direct client reqqliksense developer experience javascriptskills qliksense ja	QLIKSENSE DEVELOPER with Experience in JavaScript	$120,000
28	Preston Harr	3c8d3c73e1	https://www	New York, N	responsibilites lead work part small engineering team help ensure timely d	Senior Front End Developer	$120,000
29	JPMorgan Ch	20136a9a09	https://www	Jersey City, N	j p morgan global leader asset wealth management services serve institutic	Java Applications Developer	$120,000
30	DTG Consulti	0eb50637d6	https://www	New York, N	senior level professional capable automating portfolio management capabi	Sr. Full Stack Developer	$120,000
31	Goldman Sac	15c7fddac90	https://www	Jersey City, N	impact developer within tax operations technology team responsible devel	Java Developer - Client Tax Technology	$120,000
32	Scholastic Te	50e5bb1b42	https://www	Manhattan, I	liaison staffing looking talented software engineer java join top financial cli	Software Engineer (core Java) [Full-time / Salaried]	$120,000
33	Talentburst,	c63b578de1	https://www	New York, N	core java lead developer hands scrum experience midtown nyc primary res	Core Java Lead Developer - Scrum	$120,000
34	VLink	398ea60168	https://www	New Britain,	title software development analyst peoplesoft position type direct placem	Software Development Analyst - PeopleSoft	$120,000
35	JPMorgan Ch	556403fb4bf	https://www	Jersey City, N	jpmorgan chase co nyse jpm leading global financial services firm assets 2	Application Development Director - Control Systems Technology	$120,000
36	Bloomberg	abcb4554db	https://www	New York, N	job requisition number 56473 team trading analytics team building founda	Senior Software Engineer - Trade Analytics	$120,000
37	A+E Network	2790ac77acc	https://www	New York, N	cloud engineer responsible combination system administration platform e	Cloud Engineer	$120,000
38	TekArch	1d327bd1b9	https://www	Princeton, N	mandatory h1 f1 candidates need apply big data application architect java	Big Data Application Architect	$120,000
39	PRI Technolo	e1557e2cc5	https://www	New York, N	senior agile project managermy name bill stevens new contract hire senior	Senior Agile Project Manager	$120,000
40	Custom Grou	e816b31459	https://www	New York, N	client exciting innovative startup based new york looking full stack develop	Full Stack Developer	$120,000
41	Custom Grou	a9af791557a	https://www	New York, N	ios developerinnovated social network uniting thinkers leaders major globa	iOS Developer	$120,000
42	Open System	988f0c6ecb4	https://www	Greenwich, C	join risk technology team help build python environment support risk quan	Python Java Risk Developer	$120,000
43	Onward Sear	4418507acb	https://www	New York, N	design develop test deploy elegant software solutions client leading global	Software Engineer C++ / Python	$120,000
44	Averity	aff6474db96	https://www	New York, N	looking hire senior ruby ruby rails engineers join small back end team resp	Ruby Engineer For Tech Startup	$120,000
45	fiverr	7f5dd318d8	https://www	New York, N	fiverr looking senior backend software engineer join category experience t	Senior Backend Software Engineer	$120,000
46	BNY Mellon	0407d427ff4	https://www	New York, N	newly formed emerging business technology ebt team within investment s	Principal Blockchain Business Expert (Agile)	$120,000
47	JPMorgan Ch	461914f2991	https://www	Jersey City, N	jpmorgan chase co nyse jpm leading global financial services firm assets 2	Web GUI Developer Lead - Global Cash Portal, VP	$120,000
48	Wipro Limite	6a91e78994	https://www	New York, N	chief digital architects senior digital architects operating within wipro digit	Chief Digital Architect	$120,000

jobs.csv

II. Data Cleaning

When applying a ML method, data samples constitute the basic components. Every sample is described with several features and every feature consists of different types of values. Furthermore, knowing in advance the specific type of data being used allows the right selection of tools and techniques that can be used for their analysis. Some data-related issues refer to the quality of the data and the preprocessing steps to make them more suitable for ML. When improving the data quality, typically the quality of the resulting analysis is also improved. In addition, in order to make the raw data more suitable for further analysis, data cleaning should be performed that focus on the modification of the data. In our project, we are using job summary as a feature to fit the model. The job summary is an unstructured data having many words where each word acts as a feature. Feature reduction is done by removing all the stop words in the summary and is achieved using Natural Language Toolkit (NLTK). The Natural Language Toolkit (NLTK) is a Python package for natural language processing. NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

III. Predicting the company using machine learning algorithms

The main idea of the project is to predict getting an interview call from the company. We accomplish this by using different machine learning algorithms and also determine the best model. Here, we are training the job summary details with the company name and predicting the best company with the resume details. We extract the resume details which is in a pdf format to a text format which is used in the model to predict the company name. This can be done using pypdf2 package in python. The pypdf2 package is capable of extracting document information, splitting documents page by page, merging documents page by page etc. It allows PDF manipulation in memory and is therefore a useful tool for websites that manage or manipulate PDFs. The extracted data is again cleaned using the NLTK package to remove stopwords. Normally prediction for a text data is done depending on the frequency of word occurrence. The importance of a word is determined not by its frequency of occurrence but by its relevance. In order to increase the efficiency we assign weightage to each word in the job summary before fitting into the model. This is done by using TFIDF. Tf-idf stands for term frequency-inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. The features are assigned weights and are then used to fit the model. We used sklearn machine learning package in order to predict the company. Three different models are used:

- K Nearest Neighbor Algorithm
- Decision Tree Algorithm
- Gaussian Naïve Bayes Algorithm

1. K Nearest Neighbor Algorithm (K-NN algorithm)

K-NN is one of the simplest classification algorithm. It is a non-parametric method used for classification and regression. However, it is more widely used in classification problems in the industry. It stores all available cases and classifies new cases based on a similarity measure. K-NN has been used in statistical estimation and pattern recognition. The input consists of the k closest training examples in the feature space. The output depends on whether K-NN is used for classification or regression:

- ➢ In K-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
- ➢ In K-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The K-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

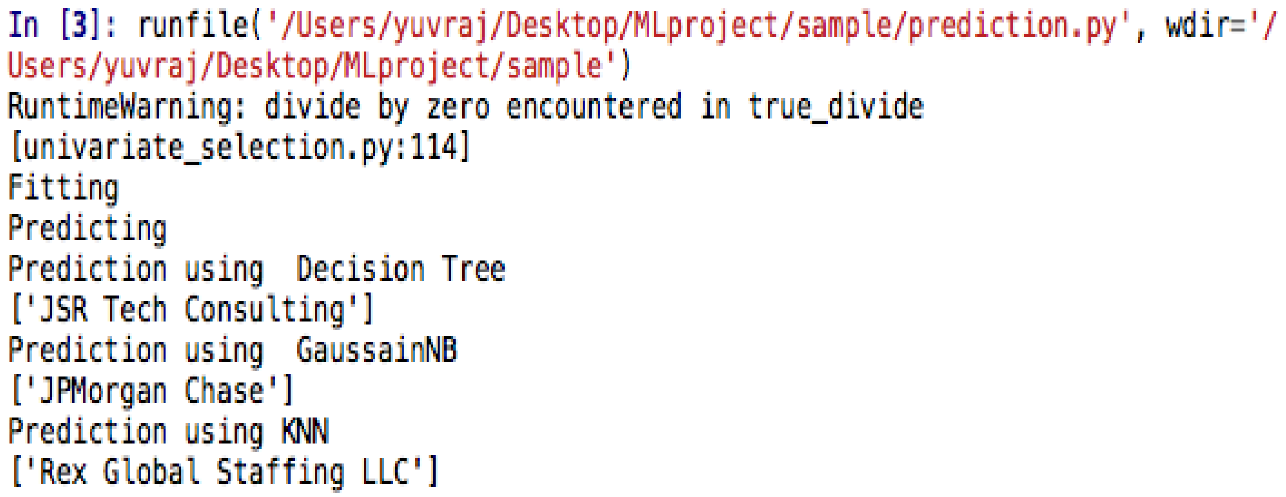
2. Decision Tree Algorithm

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables. Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

3. Gaussian Naïve Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

After fitting the features in the model we predict the company name based on resume details.

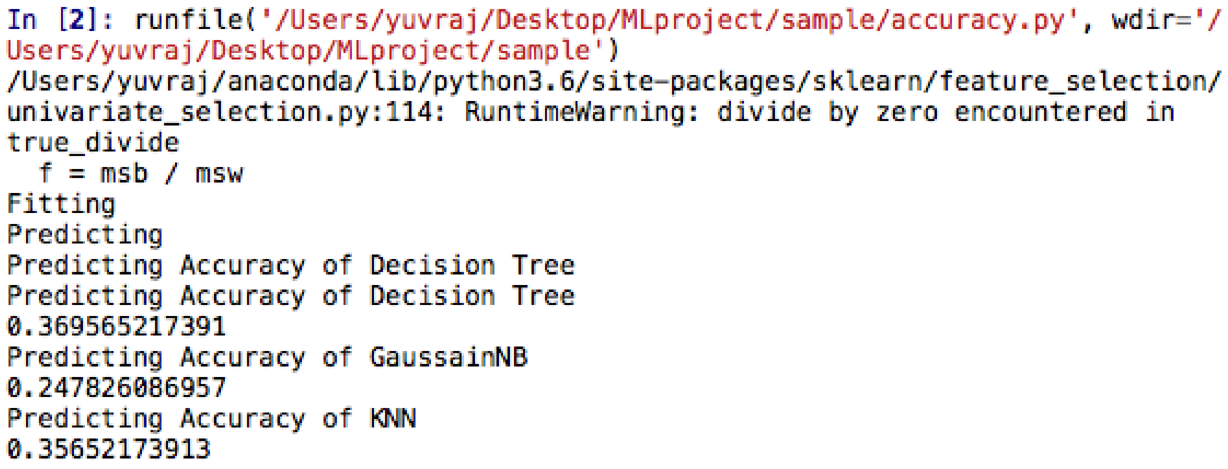


```
In [3]: runfile('/Users/yuvraj/Desktop/MLproject/sample/prediction.py', wdir='/
Users/yuvraj/Desktop/MLproject/sample')
RuntimeWarning: divide by zero encountered in true_divide
[univariate_selection.py:114]
Fitting
Predicting
Prediction using  Decision Tree
['JSR Tech Consulting']
Prediction using  GaussainNB
['JPMorgan Chase']
Prediction using KNN
['Rex Global Staffing LLC']
```

Figure shows the prediction of company name

IV. Determining the accuracy

We used sklearn package to calculate the accuracy. It is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.



```
In [2]: runfile('/Users/yuvraj/Desktop/MLproject/sample/accuracy.py', wdir='/
Users/yuvraj/Desktop/MLproject/sample')
/Users/yuvraj/anaconda/lib/python3.6/site-packages/sklearn/feature_selection/
univariate_selection.py:114: RuntimeWarning: divide by zero encountered in
true_divide
  f = msb / msw
Fitting
Predicting
Predicting Accuracy of Decision Tree
Predicting Accuracy of Decision Tree
0.369565217391
Predicting Accuracy of GaussainNB
0.247826086957
Predicting Accuracy of KNN
0.35652173913
```

Accuracy is determined by comparing the predicted value with the test value. The accuracy is less because of the unstructured data. The K nearest neighbor classification has a better accuracy than the other two and hence it is a better classification model.

Future Work

The job summary is an unstructured data. Hence the accuracy is less while predicting using unstructured data. We can increase the accuracy by converting unstructured data to structured data. By using NLTK packages we can get a more structured job summary data.

References

1. http://scikit-learn.org/stable/documentation.html
2. https://pypi.python.org/pypi/PyPDF2/1.26.0
3. https://www.crummy.com/software/BeautifulSoup/bs4/doc/
4. http://www.nltk.org/
5. http://www.tfidf.com/