Anh Nguyen

Sujata Sharma

Gurjit Singh

Machine learning can help solve many different types of business and societal problems that we face today. For example, we can look at the full time employees in San Francisco. Our group wanted to examine the data for different departments and job types and see if we can figure out what kind of departments are most similar, and if we could predict where a certain job type would fall based on its different properties.

We obtained our job classification dataset from Kaggle.com, which contains information on job titles, departments, high biweekly salary rates, low biweekly salary rates, and more. The first thing we want to do is preprocess the data. Preprocessing is an important step in machine learning because we need to work with a clean dataset that does not contain missing values or biased data. We achieved this using a couple different methods. First we checked for duplicates and got rid of those, and we also checked that job codes were unique to job titles. There were also a couple rows containing missing values, so those had to be removed as well.

In order to get a better understanding of the data, it is always a great idea to create some visualizations. Our dataset did not come with a great data dictionary, so we had to find some definitions from similar datasets and apply them to ours to help explain the department data better. One visualization we thought was pretty insightful was to plot out the mean biweekly rate for the top 20 job types in a bar graph. This helps show which kind of jobs are on the higher end of the pay spectrum. Among the highest were Chief of Police, Fire Department Chief, and Mayor. The Fire Department as a whole had the highest average pay rate. The total average biweekly high rate among all job types was 2910.65, and low rate was 3536.31. We also created visualizations for these values distributed across the dataset.

After getting more familiar with the data, we moved on to further analyze it. KNN is a type of unsupervised learning. We wanted to try out this classification method on the data to see if we can find patterns between the departments. The data was organized into the highest frequency groups, and then split into a training and test set. After running the algorithm, we can pull precision and recall values for specific departments. This can show us which departments are closely related to each other, and help us predict where a new job type would most likely fall into based on its properties. We also calculated the error for K values, and found that the mean error is lowest after K = 19.

Next we performed PCA on the raw dataset. Plotting the PCA showed positive associations with Biweekly High rate and negative associations with Biweekly Low Rate. This data was used to create hierarchical clusters. We created clusters for single link (nearest neighbor), complete link (furthest neighbor), and average link clusters. The hierarchical clusters help show us a lot of different associations between the departments. For example, Sheriff Juvenile Probation, and City and District Attorney departments are clustered together which makes sense. However, the Controller and Tax Collector departments are also in the same cluster, which is very interesting. We would not have found this association without these analytical tools. Going forward, we could try to fine tune these algorithms and test out new ones to ensure that we have the best tools available to predict and classify San Francisco’s job market.