LSE

THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE

ST451 – Bayesian Machine Learning

13917

Professor Kalogeropoulos

Department of Statistics

Lent Term 202

1.1. Introduction

The dataset of this paper is borrowed from a Kaggle competition on HR Analytics. It comes from a company that runs training courses from which it tries to hire Data Scientists, and it wants to predict which candidates are more likely to work for the company after the training. The initial dataset had 15 variables including the target variable, which is 'Yes' or 'No' depending on whether the person joined the company after undergoing the training, and 19158 observations.

Variable Name	Type (Categories)	Encoding	Number
Enrollee_Id	Categorical	Removed	/
City	Categorical (123)	Removed	/
city_development_index	Continuous	No change	/
gender	Categorical (3)	Dummy variables	4508
relevant_experience	Binary	No change	/
enrolled_university	Categorical (3)	Dummy variables	386
education_level	Categorical (5)	Dummy variables	460
major_discipline	Categorical (6)	Dummy variables	2813
experience	Discrete	Dummy variables	65
company_size	Caregorical (8)	Categorical	5938
company_type	Categorical (6)	Dummy variables	6140
last_new_job	Categorical (6)	Categorical	423
training_hours	Discrete	No change	/
target	Binary	No change	/

Table 1 - Variables

There is a large number of missing values and most of the variables are Categorical. The first part of the project consisted of a classification, for which many algorithms work only with numerical variables. If we code non-ordinal categorical variables with numbers, we would be imposing a hierarchy among the different classes that does not exist. For this reason, most categorical variables were coded as dummy variables, unless a hierarchy could be assumed without a doubt. Missing values were inferred logically where possible, sometimes relating them to other closely related variables. If not, they were from predicted on the remaining variables or removed altogether. After said transformation, we're left with a dataset of 9785 entries and 26 independent variables.

The target variable is heavily imbalanced, which makes classification challenging. However, by comparing different methods of classification in the following section and using clustering in the next one, the issue with imbalance could be tackled and It was possible to identify important variable when it comes to predict who is more likely to join the company.

2. Classification

We started doing a classification on the target variable on all the other 26 variables. We used the rather simple models of Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Logistic Regression. Other more sophisticated Bayesian Models like Gaussian Process Classification were tried but worked out much worse, especially because of the high dimensionality and amount of data made computation difficult.

The three models used yielded similar results in terms of prediction accuracy. Logistic regression has the advantage that it places less assumptions on the data and allows to do significance tests on the coefficients that determine the importance of each independent variable. However, LDA performed better when identifying positive instances of the target variable. Because of the mentioned imbalance, this is not noticeable when comparing the models in terms of prediction accuracy, but when looking at other metrics like ROC and AUC, LDA is preferred, as we will see in the results of this section.

2.1. Linear and Quadratic Discriminant Analysis

LDA tries to predict the probability of an observation belonging to one class or another and assigns that observation to the class for which the probability is highest. It makes use of the probability of an observation to belong to one class or the other without any other knowledge, also called prior probability in Bayesian statistics, and the probability distribution of the predictor variables conditional on the class they belong to. With the help of Bayes' theorem, the probability of the observation belonging to one class or the other is calculated conditionally on the predictor variables. Let's look at Bayes' Theorem in detail first to understand this:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

We can see that the probability of one event or random variable conditional on another one is equal to the probability of their intersection divided by the probability of the variable we want to condition on. From that follows that the intersection of both is equal to the conditional probability times the probability of the variable we are conditioning on. Moreover, if in a dataset we observe instances of a variable always intersecting with another one, in this case instances of the predictor variables always intersecting with the class variable, we can calculate the standalone probability of the predictor variables by:

$$P(B) = \sum_{i}^{n} P(B \cap A_i)$$

That is, the probability density of the predictor variable regardless of which class it belongs to. We use these facts in LDA and QDA to compute the conditional probability of a class as follows:

$$P(C_1|X) = \frac{P(X|C_1)P(C_1)}{\sum_i^n P(B \cap A_i)}$$

Some details worth noting is that LDA and QDA assume predictor variables to be normally distributed if they are continuous or to follow a Bernoulli or Categorical distribution if they are categorical **(Bishop, 2006)**. Also, LDA and QDA differ in that LDA assumes the predictor variables to have the same covariance matrix regardless of class, while QDA considers a different one for each class. They estimate the prior probability of a class and the required parameters for the predictor variables via maximum likelihood, the details of which we won't cover because they are beyond the scope of this work **(Bishop, 2006)**.

2.2. Logistic Regression

Logistic Regression is used for the same purposes but without making any of the said assumptions on the predictors. It assigns observations to the class of which they have a highest probability of belonging through a Bernoulli distribution.

$$Y_i \sim Bernoulli(\pi(C_k|X_i))$$

The parameter $\pi(C_1|X_i)$ of the Bernoulli Distribution is calculated with a sigmoid function on the log-odds ratio, the logarithm of the odds ratio:

$$\frac{P(X|C_1)P(C_1)}{1 - P(X|C_1)P(C_1)}$$

The idea is based on the same conditional density of a class given the predictors as above:

$$P(C_1|X) = \frac{P(X|C_1)P(C_1)}{\sum_i^n P(B \cap A_i)} = \frac{1}{1 + e^{-a}} = \sigma(a) = \sigma(\beta^T \phi)$$

The term $\frac{1}{1+ e^{-a}}$ is a sigmoid function of a and we denote it by $\sigma(a)$. Instead of a function of the log-odds ratio we can also consider it a linear function on a feature vector ϕ, which is the vector containing the predictor variables. With this approach we just have to determine the vector β. This makes model fitting much easier because we only have as many parameters to estimate as we have predictors. In LDA on the other hand we have a total of $M(M + 5)/2 + 1$ from the predictors means, covariance and priors, where M is the number of predictors.

2.3. Results

10-Fold Cross-Validated Metrics	Logistic Regression	Reduced Logistic Regression	LDA	QDA
Accuracy Rate	84.74%	83.71%	85.71%	76.41%
AUC	0.602	0.503	0.715	0.715

Table 2 - Cross-validated Model Results

Best Fold		Logistic Regression (all predictors)		LDA		QDA	
	Predicted Class	No	Yes	No	Yes	No	Yes
True Class	No	796	23	761	58	650	169
	Yes	114	45	74	85	57	102

Table 3 - Confusion Matrices per Model with 50% probability threshold for Positive Classification

At first sight, all fare relatively similar in terms of prediction accuracy, but when we look at the area under the receiver operator curve (AUC) we see that Logistic Regression is clearly inferior. The Reduced Logistic Regression is a Logistic Regression with only the predictors that were statistically significant. The problem with the low AUC in LR is the mentioned imbalance in the target variable. Because of this, positive instances of the class are hard to spot, so as soon as the model became simpler it wasn't able to accurately classify them.

LDA is the best option in terms of overall accuracy and AUC. AUC takes into account the proportion of False Positive and False Negative Errors. It is a curve that plots the True Positive as a Function of the False Positives for each probability threshold under which we would assign an observation to 1 instead of 0. The higher the area under this curve, the better we classify true instances of the target variable accurately for all possible thresholds.

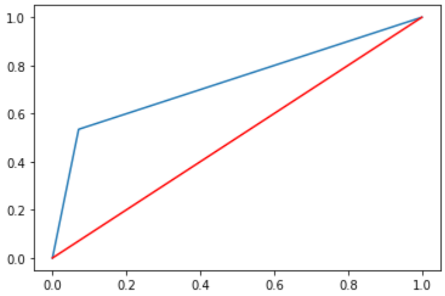

Figure 1 - LDA AUC

If we deemed one kind of error more important than the other one, for example if we wanted to reduce false negatives at the expense of increasing false positives, we could lower the threshold for classification in LDA. In the confusion matrices above we see that QDA does a classification that is more in this direction, but the better approach would be to stick to LDA and change the threshold, as it is more accurate overall. This could be a desirable option if we wanted to spot all people susceptible of joining the company at the expense of spending effort on many who are not likely to do it.

2.4. Reduced Logistic Regression and Importance of Variables

	coef	**std**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	5.8661	0.681	8.614	0.000	4.531	7.201
city_development_index	-7.6932	0.249	-30.925	0.000	-8.181	-7.206
relevent experience	-0.2295	0.090	-2.559	0.010	-0.405	-0.054
Full time course	0.1911	0.092	2.080	0.038	0.011	0.371
>20y experience	-0.7955	0.130	-6.104	0.000	-1.051	-0.540
experience	-0.0440	0.008	-5.624	0.000	-0.059	-0.029

company_size	0.0363	0.015	2.461	0.014	0.007	0.065
>4y from last job	0.2321	0.108	2.145	0.032	0.020	0.444
last_new_job	0.1151	0.034	3.390	0.001	0.049	0.182

Table 4 - Significant Variables' Coefficients and Significance Test in Logistic Regression

Despite the Logistic Regression not being the best method in terms of prediction, it allows us to draw some conclusions around which are the most important variables. We can see that negative coefficients have generally a greater magnitude than positive ones, so that makes the majority of predictions go to 0. City development index is the one that stands out in this aspect, and because its values are mostly close to one (the company looks mostly in cities with high development, the minimum values is 0.44, and the mean and median are 0.846 and 0.91, respectively) the majority of the dataset belong to the class 0. We can also see that the coefficients for experience and >20y experience are negative. Also, full time course and time past from last job have the highest positive coefficients. his points out that potentially profiles of candidates to look at are:

- Recent graduates and soon to be graduates
- People with a low to moderate amount of experience who have been in their jobs for some time (perhaps more than 2 years)

And the less economically developed the city they live in is, the higher chances they have to recruit them. These variables offer a possibility to do a rough classification. However, by removing variables from the beginning we lost the capacity to identify positives in the sample. A possible reason for this is that the company is mostly present in highly developed cities and those of lower development are of no interest, and from that point it is difficult to discriminate accurately. However, LDA offers a more precise way of classification as seen before.

3. Clustering

3.1. Gaussian Mixture Models

After the classification a cluster analysis was made on the data. The model used was Clustering with Gaussian Mixtures. This consists of a mixture of Gaussian Distributions with a latent variable that represents the cluster. We can represent the probability distribution of a given predictor conditional on the latent variable/cluster it belongs to (z) (**Bishop, 2006**) as:

$$p(x|z_k) = N(x|\mu_k, \Sigma_k)$$

z follows a categorical distribution with a prior probability of π_k. The predictor has a different mean depending on the cluster, as well as a different covariance (depending on model specifications as well). The clusters' categorical distribution is denoted by:

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

Where z_k takes on the values 0 or 1 depending on whether the observation we are dealing with belongs to the cluster or not. Based on Bayes' Theorem like in the previous section, we calculate the probability of x unconditionally of the cluster as:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

And the conditional probability of belonging to a cluster based on the predictors as:

$$P(z_k = 1\ |x) = \frac{p(z_k = 1)p(x\ |z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)}$$

The parameters are calculated iteratively with the help of the Expectation Maximization Algorithm, the details of which are beyond the scope of this work (Bishop, 2006).

3.2. Clustering Results

With a Gaussian Mixture Model, the ideal clustering was found to be 27 clusters with a diagonal covariance matrix, as it had the lowest BIC, a common metric used to choose models – a model fits the data better the lower the BIC. Some very small clusters were formed, so those with less than 100 observations were discarded. Then, we looked at those with conversion rates above the average (16.28%) to try to find specific characteristics to them and also try to validate the findings from the previous section around which variables are most important, as well as trying to find new ones.

	Selected Clusters Mean (Standard Deviation)				1st LR Coefficients	P-value	2nd LR Coefficient
Cluster Number	4	5	18	25			
Cluster Size	290	155	2106	176	/	/	/
City development index	0.78 (0.14)	0.865 (0.10)	0.76 (0.14)	0.78 (0.13)	-7.69	0.000	-7.7115
Female	0 (0)	0.28 (0.45)	0 (0)	0.89 (0.32)	-0.25	0.444	
Male	1 (0)	0.66 (0.47)	1 (0)	0.06 (0.23)	-0.25	0.422	
Relevant experience	0.87 (0.33)	0.52 (0.50)	0.8 (0.40)	0.74 (0.44)	-0.23	0.010	-0.2473
Full time course	0.15 (0.36)	0.24 (0.43)	0.26 (0.44)	0 (0)	0.19	0.038	0.2032
Part time course	0.18 (0.32)	0 (0)	0.16 (0.36)	0.22 (0.41)	-0.24	0.066	

High School	0 (0)	0 (0)	0 (0)	0 (0)	-0.30	0.556	
Graduate	0.73 (0.46)	0.23 (0.42)	0.77 (0.42)	0.76 (0.43)	-0.11	0.842	
Masters	0.25 (0.43)	0.17 (0.38)	0.23 (0.42)	0.24 (0.43)	-0.18	0.752	
Phd	0.02 (0.13)	0.6 (0.49)	0 (0)	0 (0)	0.25	0.683	
Other Major	0 (0)	0 (0)	0 (0)	0 (0)	-0.3	0.451	
Arts Major	0 (0)	0 (0)	0 (0)	0 (0)	-0.22	0.638	
Humanities Major	0.02 (0.15)	0 (0)	0 (0)	0 (0)	0.17	0.630	
Business Degree Major	0 (0)	0 (0)	0 (0)	0 (0)	0.13	0.738	
STEM Major	0.95 (0.21)	1 (0)	1 (0)	1 (0)	0.14	0.645	
>20y Experience	0 (0)	0 (0)	0 (0)	0 (0)	-0.80	0.000	-0.7337
Experience	7.45 (4.83)	9.74 (5.17)	8.10 (4.76)	8.04 (4.51)	-0.04	0.000	-0.0420
Company size	0.76 (0.85)	4.10 (2.12)	3.36 (2.19)	2.98 (2.25)	0.04	0.014	0.0419
NGO	0 (0)	0.21 (0.41)	0.10 (0.30)	0 (0)	-0.72	0.033	
Public Sector	0 (0)	0.4 (0.49)	0 (0)	0 (0)	-0.39	0.227	
Pvt_Ltd	0 (0)	0.39 (0.49)	0.81 (0.39)	0.7 (0.46)	-0.69	0.022	
Early stage startup	1 (0)	0 (0)	0 (0)	0.2 (0.4)	-0.60	0.068	
Funded startup	0 (0)	0 (0)	0.08 (0.27)	0.08 (0.27)	-0.76	0.017	
>4y from last job	0 (0)	0.17 (0.38)	0.18 (0.39)	0.22 (0.41)	0.23	0.032	0.2465
Last new job	1.29 (0.86)	1.30 (1.08)	1.24 (1.07)	1.13 (1.02)	0.12	0.001	0.1144
Training hours	73.8 (69.1)	66.2 (70.75s)	64.75 (60.78)	55.65 (47.46)	-0.0006	0.244	
Target	0.24 (0.43)	0.34 (0.47)	0.48 (0.50)	0.52 (0.50)			

Table 5 - Variables in most relevant Clusters and Linear Regression

3.3. Variable Findings

STEM Majors

One very interesting finding is that all the relevant clusters have an overwhelming majority of STEM majors, despite the coefficient in the Regression being roughly the same as that of other majors and not statistically significant.

City Development Index: opportunity to recruit highly skilled people

A confirmation of one of the previous findings is that the lower the development index of a city, the easier it is to recruit candidates. The average index for the 4 clusters is almost always below the overall mean (0.846), and median (0.91). Especially worth noting is the case of cluster 18, which is the largest of the 4 and has more than 20% of the dataset, with the lowest average city development index (0.76) and an average conversion rate of 0.48.

In 3 of the clusters ~75% of people are graduates and the remaining master's graduates. Only in cluster 5, that with the lower city development index, we see that the largest proportion are PhDs. This shows a good opportunity in recruiting in less economically buoyant areas, as it shows the possibility to recruit highly skilled people easier.

Previous Experience

Regarding experience, all the clusters have average values for years of experience on the rather lower end (below 10, which is the middle point of the range, and none of them have more than 20). This also validates the point made previously. Moreover, if we look at *relevant* experience, we see that the cluster with the least relevant experience is that with the highest city development index. We also see that cluster 5, that with the lowest city development index and the high proportion of PhDs, has the lowest number for relevant experience, strengthening the thesis that in less developed areas it is easier to recruit talent because of less competition from other employers.

Female Recruits and Frequency of Changing Jobs

Gender is a heavily imbalanced class and the findings about that class from clustering are noteworthy. In a dataset of 9785 instances, only 831 are Female (8.5%). 137 of them converted, that is, 16.5% of females converted, but they represented only 8.6% of conversions (1593). On the other hand, of the 8857 males, 1442 converted, so the conversion rate is very similar at 16.3%, but these represent the vast majority of conversions (90.5%, as some people are classified with 'Other' as gender). The two clusters that contain women have ~200, so it is most likely that all the 137 that converted are included in them. The class imbalance, however, probably causes the Logistic Regression coefficients to be roughly the same so we cannot draw any conclusions or build any profiles based on gender.

If we look at cluster 25, which is comprised to 89% of women, we can see that all of them have STEM majors, like the vast majority of other conversions. One difference to the rest, though, is that this cluster has ~28% of members coming from startups, unlike in the other 3, which have none or just 8%. This is also the cluster with the highest proportion of people who haven't switched jobs in the last 4 years (22%), which is consistent with the finding from the previous section that people are more likely to join the company the longer they have been at their present job. When checking women who are in Early-Stage Startups, we see that none of them has been there for longer than 4 years, and only one was in the case of Funded startups, so all women who converted and had been with their employer for longer than 4 years must come from normal Private Limited Companies.

These findings are particularly interesting for diversity purposes. If the company were looking to hire more women, they would know they should have to target those who have been for longer times with their employers, and those who are currently at startups.

Enrollment and Company Size

We see generally values between 15-25% for either Full- or Part-time enrollment. Although this might seem low, the proportion in the overall population is 10.4% for Full time and 6.3% for Part time, so the members of these clusters where people are more likely to join the company are also more likely to be enrolled in education. With regards to company size, most clusters have larger numbers, according to the statistically significant but relatively low regression coefficient for that variable.

4. Conclusion

After analyzing the data with classification and clustering procedures, we found out that the most important variables were:

- City Development Index
- Relevant Experience
- Full time course enrollment
- Years of Experience
- Company size
- Time Spent in the last Job
- STEM Major

The most important variable seems to be the City Development Index, and that is probably to be attributed to a more competitive labor market in the cities with a higher development. When looking at the selected cluster with the lowest index we saw that it was more frequent to find people with higher academic classifications, but that they had the lowest proportion of relevant experience, due to lack of similar opportunities. Because of this we can conclude that less developed cities are good pools for finding talent.

Also, we saw that there is a low portion of women in the dataset, but despite this they have the same rate of conversion as men. Among women, we found slightly different characteristics, like a higher proportion of the in startups, and a larger proportion of them having spent more than 4 years in their current job if they work for a Private Limited Company, so in case the company wanted to hire more women for diversity purposes, these would be good potential pools.

Bibliography

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning.* s.l.:Springer.