STSCI 4740 Final Project Report

Group Member: Yiting Zhong, Shiyun Wang, Yiran Wang, Aria Gao

1. Introduction

Job seekers never stop equipping themselves with the skills and attributes preferred by employers. A clear and concise summary of the job requirements for different job positions should be very useful. In this project, with the given data set containing 1200+ Google jobs description, we constructed Decision Tree, Random Forest, Naïve Bayes, Logistic Regression and Association Rule Mining models to identify important features for particular job categories. The final evaluation of classification models' performances is displayed with two metrics, Confusion Matrix and ROC Curve. In this report, we would initiate the step-by-step process from the beginning.

2. Keyword Extraction

The dataset "job_skill_short.csv" contains only unstructured text data of phrases and paragraphs. In order to detect important job characteristics, we split sentences in Responsibilities, Minimum.Qualifications and Preferred.Qualifications into individual words and count their frequencies. For these three columns, we took five steps to extract important keywords:

i. Separated words by blanks;
ii. Used Regular expressions to keep only letters and numerals with other punctuations replaced by blank;
iii. Removed extra blanks and null values;
iv. Filtered meaningless common words, such as 'a', 'the', etc., using the stop-word list[1], (Silva 1662);
v. Obtained final words and their frequencies and kept top 100 frequent keywords.

2.1. Automatic Feature Generation

Noticed that not all top 100 frequent words were helpful in identifying a job category. For example, 'degree' is the most frequent word for minimum qualifications, but it cannot necessarily differentiate different jobs. Among the keywords obtained, we further validated meaningful features based on business common sense of job-related characteristics. In consideration of potential high-dimensional problem, we integrated these words from the three columns to count overall frequencies and kept the most frequent 30 words as features in the final data set.

The 30 features generated are:

Management, Business, Technical, Product, Strategy, Team, Customer, Sales, Cloud, Marketing, Data, Engineering, Design, Communication, Partner, Solutions, Project, Cross-functional,

[1] A stop-word list is built with words that are non-informative and that are supposed to be filtered in the document representation process.

Analytical, Research, English, Write, Science, Interpersonal, Organizational, Platform, Consulting, Mobile, Fast-paced, Web.

2.2. Special Feature Generation

There are four features that are defined and coded, shown as follows:

Year_Requirement: Use preferred year requirement if it exists, otherwise use minimum year requirement. Includes 542 missing values. A continuous variable.
Degree_Requirement: Use preferred degree requirement if it exists, otherwise use minimum degree requirement. Includes 79 missing values. A discrete variable.
Title_Level: Three-class variable with values 1,2,3. 1 denotes Intern; 2 denotes entry level positions, including Analyst, consultant, Senior, Specialist and Unknown; 3 denotes management level positions, including Director, Executive, Head, Lead, and Manager. A discrete variable.
Programming_Language: Dummy variable with values 1 and 0. 1 denotes the following words appear in the job description at least one time: Python, java, C++, SQL, SAS.

We noticed that the missing rate for Year_Requirement is as high as 44.7%, meaning we could lose nearly half the data if this feature was incorporated into the models. Our preliminary experiments using data with and without Year_Requirement indicates that this feature does not have a significant contribution to classification, and thus we left this predictor out in later analysis.

3. Classification Modeling

Before models were constructed, it is noticeable that the predicted variable Y, here job category, had too many levels (23 levels). In order to ensure sufficient observations for each classification label, we divided job categories into two main classes: technical fields and non-technical fields. Non-technical fields contain the following 7 job categories: “Administrative”, “Business Strategy”, “Legal & Government Relations”, “Marketing & Communications”, “Partnerships”, “People Operations”, and “Real Estate & Workplace Services”. The remaining 16 categories were thus labeled as technical fields. We set it as a dummy variable, with 1 denoting technical field and 0 denoting non-technical fields.

3.1. Decision Tree Model

Decision Tree model is commonly used for classification problem with categorical variables as predictors. In model construction phase, we used 10-fold cross-validation method to select the best model while controlling the minimum sample size for each splitting group to be no less than 10. Figure1 shows how data is split step by step.

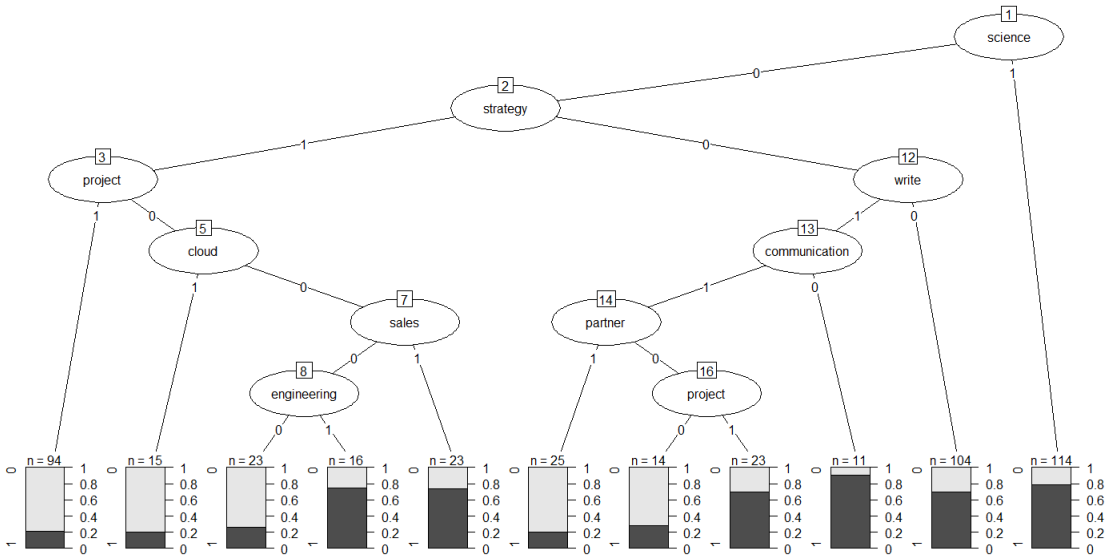

Figure 1 Unpruned Decision Tree Model Analysis

The shaded proportions of the bars on the bottom represent the probability of Y being classified as technical fields. In the figure above, three leaves with largest sample sizes—94, 104 and 114 respectively— indicate that the corresponding splitting variables— Science, Strategy, Project, and Write are the most important attributes.

However, the unpruned tree model seems to be too complicated and may lead to the overfitting problem. Therefore, we tried to prune the existing tree model based on CP (Cost Complexity) parameter, which is used to control the size of the decision tree and to select the optimal tree size by setting a cut point value to determine whether to continue the tree construction. From Appendix Figure A-2 we chose CP=0.02 since the corresponding cross-validation is relatively small. The plot of pruned tree model is shown as follows.

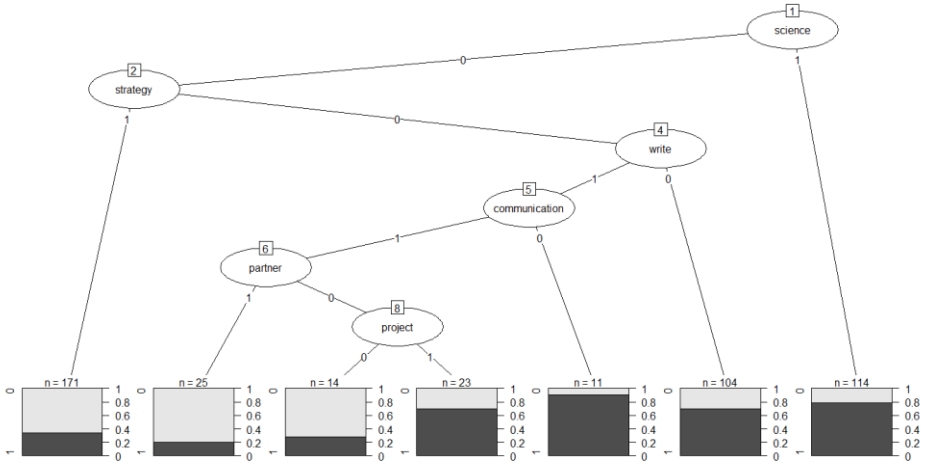

Figure 2 Pruned Decision Tree Model Analysis

In the pruned tree model, only 6 features are kept, and they are Science, Strategy, Write, Communication, Partner and Project.

3.2. Random Forest Model

The Random Forest Model provides an improvement over Decision Tree Model by way of small tweak that decorrelates the tree, using bootstrapped training samples. For the Random Forest model, we determined the number of split attributes k and decision tree number n through the following two steps, (Bannerman-Thompson et al. 107):

i. Let k ranging from 1 to 35 (number of all features). Then, we built 35 corresponding models with parameter of tree numbers set at 1000, and plotted test error rate for each model. Set k to be the one that generates lowest test error rate. After the experiment, we got the optimal k=4.

ii. Built random forest models on k=4 and plotted the number of trees versus test error rate. From Figure 3, we can clearly see that the curve of test error rate starts to go smoothly and reaches the minimum. Therefore, we chose the optimal n=100.

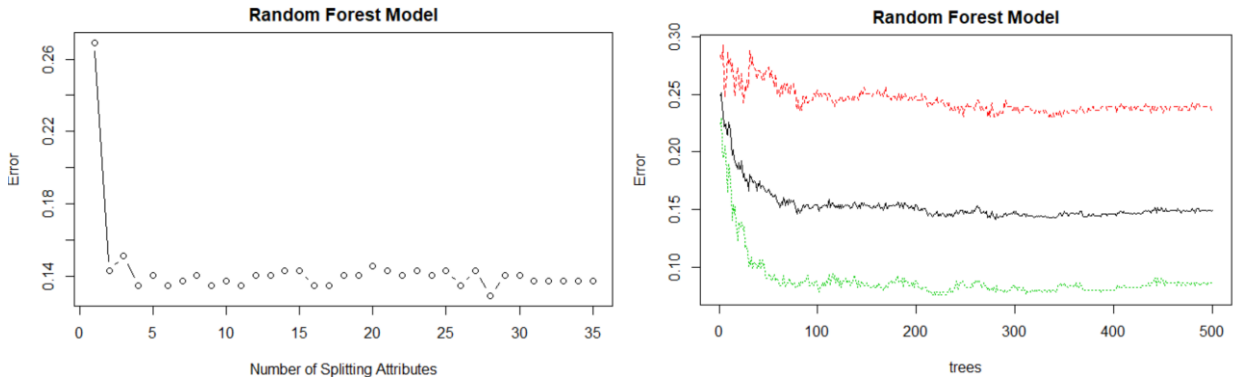

Figure 3 Number of Splitting Attributes & Trees Versus Test Error Rate

Plugging in k=4 and n=100, we obtained the final Random Forest model, and through the importance ordering shown in Appendix Figure A-3, we discovered that Degree, Job Title, Science and Communication are the most important features in classifying job categories.

3.3. Naïve Bayes Model

Naïve Bayes Model gives probabilities for different predictors in condition of job category being technical fields or non-technical fields. By printing out these conditional probabilities (as shown in Appendix Figure A-4), we could conclude that the probabilities that certain features appear for different categories have large discrepancies. These features include: Title_Level, Programming_Language, Product, Engineering, Communication, English, Science, Platform, and Consulting. In other words, they are potential important characteristics of dividing job categories.

3.4. Logistic Regression Model

We also used Logistic Regression Model for this two-class classification problem because Logistic Regression model offers significance level for all predictors in the model, which are the indicators of how significantly the predictors affect the outcome. In addition, most predictors have values between 0 and 1, so their parameter estimates are straightforwardly comparable in the same scale. Table 1 shows predictors whose p-values are below 0.1. According to p-values here, we selected the most significant features and they are Communication, Sales, Marketing, Engineering, Partner, English and Consulting. With positive parameter estimates, Sales, Engineering and English are more related to technical job categories. We also notice that having large absolute parameter estimation value, Communication and English play an important role in dividing job category groups from the perspective of practical significance.

Table 1 Significant Features for Logistic Regression Model

Feature Name	Estimate	Standard Error	P-Value	Significance Level[2]
Programming Language	0.563	0.325	0.083	.
Customer	0.395	0.214	0.066	.
Sales	0.656	0.237	0.006	**
Marketing	-0.659	0.220	0.003	**
data	-0.488	0.198	0.013	*
Engineering	0.658	0.220	0.003	**
Design	0.404	0.226	0.074	.
Communication	-0.972	0.205	0.000	***
Partner	-0.659	0.207	0.001	**
Analytical	0.380	0.199	0.056	.
Research	0.614	0.284	0.030	*
English	0.858	0.263	0.001	**
Science	0.631	0.302	0.037	*
Platform	0.465	0.268	0.083	.
Consulting	-0.600	0.230	0.009	**

Since many features, as shown in Table 1, are not statistically significant in the model, some of them may be dropped based on certain rules.

First, we used backward stepwise selection method to select a subset of useful predictors. The summary of the reduced model is shown in Table A-1. After variable selection, the number of predictors was reduced from 36 to 19. Even better, p-value of most remaining predictors decreased. Hence, the overall reduced Logistic Regression model is more concise and accurate than the original one.

Lasso Logistic Regression Model was also constructed to make a further feature selection. Using λ_{lse} as the optimal tuning parameter, we got a sparse model with coefficient estimates of 11 predictors set to zero. Detailed information about Lasso Model is shown in Appendix Table A-2.

[2] Significance Nodes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

3.5. Classification Model Comparison

3.5.1. Confusion Matrix

To see prediction accuracy of different classification algorithms, here we compared original Decision Tree and Logistic Regression Model by looking at the confusion matrix. A 2*2 confusion matrices were set up and metrics including sensitivity, accuracy, specificity and Mcnemar's p-value were calculated and displayed here to compare prediction performances between these four classification models. Results in Table 2 indicate that all the four models have decent predictions about job categories. Random Forest Model has the best performance with the accuracy of 85%, and the other three models have test error rate around 30%. Naïve Bayes Model seems to be quite robust in that its sensitivity, accuracy and specificity are very close to each other. Logistic Regression did worst in categorizing true technical fields jobs. Mcnemar's p-value is generated from ${z_0}^2$ which measures the consistency of classification models, and small p-values of Decision Tree Model and Random Forest Model tell us that these two models are relatively less consistent at doing correct predictions.

$$Accuracy = \frac{TP}{TP+FN} \tag{3.1}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3.2}$$

$$Specificity = \frac{TN}{TN+FP} \tag{3.3}$$

$$\text{Under } H_0\text{: } \pi_{1+} = \pi_{+1}, {z_0}^2 = \frac{(FN-FP)^2}{FN+FP} \sim {\chi^2}_{(1)} \tag{3.4}$$

Table 2 Confusion Matrix Metrics for Different Classification Models

	Sensitivity	Accuracy	Error Rate	Specificity	Mcnemar's P-Value
Decision Tree	0.614	0.739	0.261	0.829	0.024
Random Forest	0.778	0.854	0.146	0.910	0.054
Naïve Bayes	0.686	0.687	0.313	0.687	0.111
Logistics Regression	0.594	0.712	0.288	0.796	0.079

3.5.2. ROC Curve

In order to get a more intuitive understanding of prediction performances of these models, we generated ROC curves for models in a graph. Figure 4 displayed 4 ROC curves with different colors representing different models. By comparing AUC value, it is evident that Random Forest Model performs much better than other models. However, it could be rather vague to tell which of the rest three models performs better. Another noteworthy phenomenon is that the ROC curve of Decision Tree Model is the most smoothing one, corresponding to the fact that it predicts the same value (probability) for samples which belong to the same leaf.

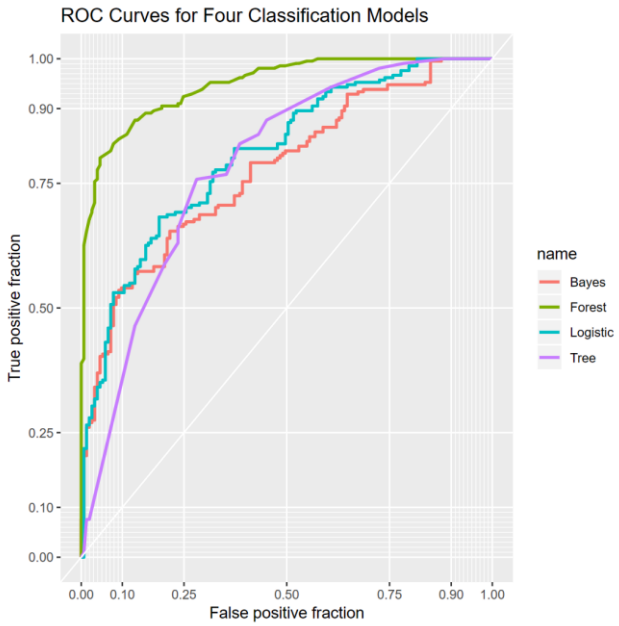

Figure 4 ROC Curves for Different Classification Models

4. Association Rule Mining Modeling

Apart from common classification models, we also tried to build association rule mining models to see the correlation between features and job categories from another perspective. In order to explore those relationships, we particularly set the value of 30 extracted features to be 1 in the left-hand side conditions, for instance, "business = 1". And we correspondingly set the right-hand side conditions to be "category_group = 1" and "category_group = 0" and built two association rule models respectively.

When fitting the model, we set the minimum Support threshold to be 0.005 and the minimum Confidence threshold to be 0.8, and then displayed only first rules with 10 largest Support to focus on most significant features. We evaluated effectiveness of the rules generated by Lift, which is a measure of how two conditions in two sides are correlated. A large Lift value above 1 indicates that the conditions are strongly positively related.

```
     lhs                                          rhs           support   confidence lift     count
[1]  {science=1}                               => {cat_group=1} 0.2128713 0.8431373  1.443337 258
[2]  {technical=1,engineering=1}               => {cat_group=1} 0.2070957 0.8175896  1.399602 251
[3]  {team=1,customer=1,sales=1}               => {cat_group=1} 0.1938944 0.8362989  1.431630 235
[4]  {Title_Level=2,technical=1}               => {cat_group=1} 0.1872937 0.8376384  1.433923 227
[5]  {Management=1,technical=1,engineering=1}  => {cat_group=1} 0.1848185 0.8086643  1.384324 224
[6]  {Title_Level=2,engineering=1}             => {cat_group=1} 0.1790429 0.8250951  1.412451 217
[7]  {Management=1,team=1,customer=1,sales=1}  => {cat_group=1} 0.1790429 0.8443580  1.445426 217
[8]  {Management=1,science=1}                  => {cat_group=1} 0.1740924 0.8683128  1.486434 211
[9]  {engineering=1,science=1}                 => {cat_group=1} 0.1707921 0.8589212  1.470357 207
[10] {product=1,science=1}                     => {cat_group=1} 0.1699670 0.8691983  1.487950 206
```

Figure 5 Association Rules on Technical Fields Job Categories

We can see from Figure 5 that the corresponding Lift values for these 10 rules are all above 1, and that Science, Technical, Team, Management and entry level job positions appear most frequently together with technical job category groups.

	lhs		rhs	support	confidence	lift	count
[1]	{Management=1,business=1,interpersonal=1,organizational=1}	=>	{cat_group=0}	0.06765677	0.8200000	1.971905	82
[2]	{project=1,interpersonal=1,organizational=1}	=>	{cat_group=0}	0.06188119	0.8152174	1.960404	75
[3]	{Management=1,project=1,interpersonal=1,organizational=1}	=>	{cat_group=0}	0.06105611	0.8131868	1.955521	74
[4]	{Title_Level=1,Management=1,business=1}	=>	{cat_group=0}	0.05940594	0.8780488	2.111498	72
[5]	{business=1,project=1,interpersonal=1,organizational=1}	=>	{cat_group=0}	0.05858086	0.8554217	2.057085	71
[6]	{Degree_Requirement=MBA}	=>	{cat_group=0}	0.05775578	1.0000000	2.404762	70
[7]	{Degree_Requirement=MBA,business=1}	=>	{cat_group=0}	0.05775578	1.0000000	2.404762	70
[8]	{Title_Level=1,business=1,project=1}	=>	{cat_group=0}	0.05775578	0.8860759	2.130802	70
[9]	{Title_Level=1,Management=1,business=1,project=1}	=>	{cat_group=0}	0.05775578	0.8860759	2.130802	70
[10]	{Management=1,business=1,project=1,interpersonal=1,organizational=1}	=>	{cat_group=0}	0.05775578	0.8536585	2.052846	70

Figure 6 Association Rules on Non-Technical Fields Job Categories

We can see from Figure 6 that that the corresponding Lift values for these 10 rules involving non-technical fields are even larger. We have more confidence to say that Project, Interpersonal, Organizational, Management and Business appear most frequently together with non-technical job category groups.

5. Conclusion

In this project, we used several machine learning methods to analyze the important attributes and skills in 1200+ google job descriptions. By applying text splitting methods, we accomplished the feature selection and generated 35 high-frequency keywords as our predictors. In order to ensure enough observations for each classification label, we divided job categories into two main classes: technical fields and non-technical fields. Then, we used Decision Tree, Random Forest, Naïve Bayes, and Logistic Regression classification models to see how significantly each predictor affect the classification of job categories. After that, Confusion Matrix and ROC Curve were used to compare the prediction accuracy of four classification models. Confusion Matrix told us that Random Forest Model has the best performance with the accuracy of 85%, and the other three models have test error rate around 30%. Then by looking at the ROC curves of the four models, we further verified the effectiveness of Random Forest, and found that the ROC curve of Decision Tree Model is the most smoothing one. Finally, Association Rule Mining models were added to detect the relationship between different features and job categories through metrics of Support, Confidence and Lift.

Based on our analysis above, we can make some implications on which features are relatively more important for certain job categories. In general, Communication, English, Science, Partner, Consulting, and Title Level are critical in dividing technical and non-technical job categories. We can conclude that non-technical job candidates could often have higher job level, but are expected by Google to have strong communication skills as well as the ability to cooperate well with partners. Technical job candidates, on the other side, besides the specific science and engineering skills, are required to have decent soft skills such as the good interpersonal and communication skills, etc.

References

Bannerman-Thompson, Hansen, et al. “Bagging, Boosting, and Random Forests Using R.” *Handbook of Statistics - Machine Learning: Theory and Applications Handbook of Statistics*, vol. 31, 17 May 2013, pp. 101–149., doi:10.1016/b978-0-444-53859-8.00005-9.

Silva, C., & Ribeiro, B. (n.d.). The importance of stop word removal on recall values in text categorization. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, pp. 1661-1666., doi:10.1109/ijcnn.2003.1223656

Appendix

Table A-1 Top 10 Frequent Features for 23 Job Categories

Job Category	Common Frequent Features	Special Frequent Features
Program Management		Marketing
Manufacturing & Supply Chain	Business	Marketing
Technical Solutions		Engineering
Developer Relations	Technical	Engineering
Hardware Engineering		Engineering
Partnerships	Management	Partner
Product & Customer Support		Engineering
Software Engineering	Product	Engineering
Data Center & Network		Engineering
Business Strategy	Cloud	Engineering
Technical Writing		Engineering
Technical Infrastructure	Sales	Engineering
IT & Data Management		Engineering
Marketing & Communications	Data	Engineering
Network Engineering		Engineering
Sales & Account Management	Customer	Engineering
Sales Operations		Engineering
Finance	Team	Engineering
Legal & Government Relations		Marketing
Administrative		Marketing
User Experience & Design		Marketing
People Operations		Marketing
Real Estate & Workplace Services		Marketing

Table A-2 Lasso Sparse Logistic Regression Model Coefficient Estimates

Predictor	Estimate	Predictor	Estimate
Degree_Requirement_Bachelor	.	Cloud	.
Degree_Requirement_Master	0.012	Marketing	-0.252
Degree_Requirement_MBA	-2.596	Data	-0.083
Degree_Requirement_PhD	.	Engineering	0.487
Title_Level2	0.304	Design	0.175
Title_Level3	.	Communication	-0.624
Pogramming_Language	0.364	Partner	-0.282
Management	.	Solutions	.
Business	-0.044	Project	.
Technical	0.039	Crossfunctional	.
Product	.	Analytical	0.073
Strategy	-0.012	Research	0.066
Team	.	English	0.569
Customer	0.131	Write	.
Sales	0.197	Science	0.659
Consulting	-0.337	Interpersonal	.
Mobile	0.233	Organizational	.
Fastpaced	.	Platform	0.194
Web	.		

Figure A-1 Wordclouds for Responsibilities, Minimum & Preferred Requirements

```
Classification tree:
rpart(formula = cat_group ~ ., data = training[, -2], control = ct)

Variables actually used in tree construction:
 [1] analytical         communication      Degree_Requirement design            engineering
 [6] marketing          partner            sales              science           strategy
[11] technical          Title_Level        write

Root node error: 351/848 = 0.41392

n= 848

        CP nsplit rel error  xerror     xstd
1 0.074074      0   1.00000 1.00000 0.040863
2 0.071225      2   0.85185 0.95157 0.040537
3 0.025641      4   0.70940 0.85470 0.039669
4 0.019943      6   0.65812 0.79772 0.039016
5 0.018519     11   0.55556 0.78348 0.038836
6 0.017094     13   0.51852 0.76353 0.038572
7 0.014245     14   0.50142 0.72934 0.038087
8 0.011396     15   0.48718 0.74074 0.038253
9 0.010000     17   0.46439 0.73504 0.038171
```

Figure A-2 Cost Complexity for Tree Model at Different Splits

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Degree_Requirement	9.412366	10.174041	11.538867	25.386321
Title_Level	10.538611	10.418888	12.894886	22.112844
engineering	9.966010	10.744025	11.757530	17.342823
science	8.800015	8.326451	10.582629	17.023172
communication	7.917283	10.513710	11.691866	13.924615
sales	9.668561	10.754637	12.582881	13.667729
project	8.916128	8.009939	10.636468	12.467027
consulting	8.561534	8.694913	9.973615	12.228737
english	6.627926	10.993519	11.347516	11.972533
partner	6.743654	9.414309	10.302740	11.935696
strategy	9.759814	8.094884	11.973831	11.818931
analytical	7.173961	8.258344	10.164347	11.235320
Pogramming_Language	8.095870	7.523781	9.561428	11.205012
Management	7.944524	7.766955	10.069430	11.014717
crossfunctional	7.120733	7.826951	9.807509	10.871586
customer	5.927123	8.234617	9.897573	10.864001
marketing	6.811513	7.835327	9.731182	10.811976
write	6.860107	8.930269	11.495889	10.761141
technical	7.750040	8.234273	9.065083	10.712324
product	6.488774	5.570919	7.506633	10.423564
data	8.298447	7.538603	10.253108	10.393861
design	6.424085	8.076504	8.810857	10.385556
business	8.101291	7.825385	11.041332	10.176656
interpersonal	5.808928	9.032910	10.067190	9.789380
team	8.263894	5.759379	9.949019	8.991737
fastpaced	6.409924	5.634680	8.180308	8.903246
platform	6.821335	6.470755	8.032603	8.536774
solutions	6.271140	6.708426	8.284155	7.774217
organizational	5.336216	7.337339	8.264441	7.470668
cloud	6.077552	6.307184	7.560713	7.128598
research	2.117581	5.628643	5.456005	5.520504
web	3.384132	5.701955	6.592885	4.905119
mobile	4.826519	5.071878	6.109657	4.891464

Figure A-3 Random Forest Model Importance of Features, Sorted by AVG Gini Index Decrease

```
Conditional probabilities:
   Degree_Requirement
Y                   Bachelor      Master         MBA          PhD
  0 0.042735043 0.732193732 0.068376068 0.148148148 0.008547009
  1 0.072434608 0.704225352 0.207243461 0.000000000 0.016096579

   Title_Level
Y           1         2         3
  0 0.2136752 0.2820513 0.5042735
  1 0.0945674 0.5211268 0.3843058

   Pogramming_Language
Y            0          1
  0 0.92307692 0.07692308
  1 0.71629779 0.28370221

   Management
Y          0         1
  0 0.1823362 0.8176638
  1 0.2032193 0.7967807

   business
Y          0         1
  0 0.2364672 0.7635328
  1 0.3360161 0.6639839

   technical
Y          0         1
  0 0.6666667 0.3333333
  1 0.5110664 0.4889336

   product
Y          0         1
  0 0.4928775 0.5071225
  1 0.3279678 0.6720322

   analytical
Y          0         1
  0 0.5698006 0.4301994
  1 0.5835010 0.4164990

   research
Y          0         1
  0 0.8917379 0.1082621
  1 0.8470825 0.1529175

   english
Y          0         1
  0 0.8319088 0.1680912
  1 0.6760563 0.3239437

   write
Y          0         1
  0 0.5441595 0.4558405
  1 0.5251509 0.4748491

   science
Y           0          1
  0 0.91168091 0.08831909
  1 0.64587525 0.35412475

   interpersonal
Y          0         1
  0 0.6951567 0.3048433
  1 0.8088531 0.1911469

   organizational
Y          0         1
  0 0.7150997 0.2849003
  1 0.8249497 0.1750503

   strategy
Y          0         1
  0 0.5783476 0.4216524
  1 0.6317907 0.3682093

   team
Y          0         1
  0 0.2393162 0.7606838
  1 0.2173038 0.7826962

   customer
Y          0         1
  0 0.6068376 0.3931624
  1 0.5191147 0.4808853

   sales
Y          0         1
  0 0.6068376 0.3931624
  1 0.5291751 0.4708249

   cloud
Y          0         1
  0 0.8176638 0.1823362
  1 0.7364185 0.2635815

   marketing
Y          0         1
  0 0.4814815 0.5185185
  1 0.5573441 0.4426559

   data
Y          0         1
  0 0.6353276 0.3646724
  1 0.5915493 0.4084507

   engineering
Y          0         1
  0 0.7720798 0.2279202
  1 0.5050302 0.4949698

   design
Y          0         1
  0 0.7863248 0.2136752
  1 0.6418511 0.3581489

   communication
Y          0         1
  0 0.3304843 0.6695157
  1 0.5090543 0.4909457

   partner
Y          0         1
  0 0.5071225 0.4928775
  1 0.5311871 0.4688129

   solutions
Y          0         1
  0 0.7635328 0.2364672
  1 0.6398390 0.3601610

   project
Y          0         1
  0 0.3903134 0.6096866
  1 0.4869215 0.5130785

   crossfunctional
Y          0         1
  0 0.6410256 0.3589744
  1 0.6297787 0.3702213

   platform
Y          0         1
  0 0.8547009 0.1452991
  1 0.6820926 0.3179074

   consulting
Y          0         1
  0 0.6752137 0.3247863
  1 0.8350101 0.1649899

   mobile
Y          0          1
  0 0.93732194 0.06267806
  1 0.84507042 0.15492958

   fastpaced
Y          0         1
  0 0.8233618 0.1766382
  1 0.8008048 0.1991952

   web
Y          0          1
  0 0.91452991 0.08547009
  1 0.81287726 0.18712274
```

Figure A-4 Naïve Bayes Model Conditional Probabilities for Each Predictor

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.3730     0.4333   0.861 0.389318
Degree_RequirementBachelor -0.1956     0.3663  -0.534 0.593340
Degree_RequirementMaster    0.1780     0.4597   0.387 0.698584
Degree_RequirementMBA     -18.0558   517.1599  -0.035 0.972149
Degree_RequirementPhD      -0.9111     0.8758  -1.040 0.298186
Title_Level2                0.2198     0.3524   0.624 0.532772
Title_Level3               -0.2544     0.3604  -0.706 0.480231
Pogramming_Language1        0.6439     0.2968   2.170 0.030040 *
Management1                 0.3560     0.2431   1.464 0.143104
customer1                   0.3859     0.2031   1.900 0.057441 .
sales1                      0.6234     0.2310   2.699 0.006948 **
marketing1                 -0.6991     0.2114  -3.307 0.000945 ***
data1                      -0.5332     0.1884  -2.830 0.004659 **
engineering1                0.6522     0.2071   3.150 0.001634 **
design1                     0.4260     0.2202   1.934 0.053095 .
communication1             -0.9941     0.1927  -5.159 2.48e-07 ***
partner1                   -0.6880     0.1992  -3.455 0.000551 ***
crossfunctional1            0.3024     0.1892   1.598 0.110009
analytical1                 0.3536     0.1829   1.934 0.053162 .
research1                   0.5561     0.2702   2.058 0.039629 *
english1                    0.7872     0.2107   3.736 0.000187 ***
science1                    0.5987     0.2882   2.077 0.037774 *
platform1                   0.3803     0.2382   1.597 0.110255
consulting1                -0.6121     0.2185  -2.801 0.005091 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1150.32  on 847  degrees of freedom
Residual deviance:  842.86  on 824  degrees of freedom
AIC: 890.86

Number of Fisher Scoring iterations: 16
```

Figure A-5 Logistic Regression Model after Backward Selection