Naive Bayes Classifier: Predicting Job Completion Time

Researched by:
Hayley Haught, Carolyn Heard, Kaylan Johnson, and Sarah Young

Abstract

As part of our project, we partnered with MasterScapes, an Abilene-based landscape company, to perform an analysis on the data that the company has collected from January to October 2015. We were instructed by our client to find the difference between the budgeted and actual hours performed, and then analyze the difference according to "service type," "crew," "month," and "zip code." After organizing the data, deleting missing variables, and partitioning the data, we were able to run Naive Bayes Classifier.

The results from this analysis showed us that MasterScapes does a great job of accurately estimating the amount of time it will take to complete a job. In fact, 97% of the jobs are completed within the budgeted time. While this is great information for the company, we have included some suggestions that we believe will help MasterScapes improve their service in the future. By looking at the results of our analysis, we suggest that MasterScapes budget more hours for the months of March, April, and May, seeing as these are the busiest months. We also suggest that the company focus on improving the efficiency of one of the crews in Abilene and the only Sweetwater crew, as these are the slowest. These are just a couple of the suggestions we believe will help MasterScapes; a detailed description of all of our suggestions is included in this report.

Introduction

We gathered our data from a local business called MasterScapes. They are an Abilene-based landscape company that focuses on design, maintenance, chemical treatments, and irrigation. We

received an Excel file containing raw data that provided us with information on all jobs completed through the month of October of this year. The information includes client zip codes, service types, budgeted hours, actual hours, assigned work crews, and dates completed.

We were asked to first find the difference between budgeted and actual hours and then analyze the difference according to "service type", "crew", and "month". This analysis will give MasterScapes a better understanding of how accurately they budget jobs. In our analysis, they are also able to see which work crews consistently finish jobs on time, which service types take the longest to complete, and which months are the slowest in terms of business.

Once we had the difference in budgeted and actual hours, we were asked to create a pivot table with the zip codes of under-budgeted jobs. This will tell MasterScapes if there are certain zip codes that get under-budgeted more often than others. It is important for managers to know if a certain location consistently takes longer to complete so that they can speed up the job process. The estimation of budgeted hours is especially critical for the bidding process. MasterScapes must know within reason their ability to meet the completion date agreed upon by the client. Failure to accurately estimate budgeted hours can obstruct the bidding process and ultimately result in loss of future sales. Therefore, budgeting accuracy is essential to success.

In an article on estimating job flow times, author Bernard Taylor wrote:

> "It is noted that the estimation of job flow time to meet delivery promises is extremely difficult and depends not only on the characteristics of the individual job and the particular dispatching rule, but also on the nature and status of other

> jobs in process during the same time period that the job under consideration will be present in the job shop."

In MasterScapes' case, predictors for under- or over-budgeting are the month in which the job is completed, service type, and assigned crew. Ideally, the information gathered from our analysis will help MasterScapes improve time estimations for clients, increase efficiency within work crews, and make better scheduling decisions in the future.

Methodology

For this project, Hayley Haught was the liaison between our group and MasterScapes, as her parents are the owners. We obtained our data from Andrew Moehlman, one of the department managers and the primary business analyst for MasterScapes. His responsibilities are focused around developing and evaluating metrics for all departments. Andrew sent Hayley an email containing raw data exported as an Excel file from the company's customer relationship management software, Service Autopilot. This data detailed each job completed by the MasterScapes crews dated from January to October of 2015. Andrew explained that he wanted to see a difference between budgeted hours and recorded hours in order to determine how well the account managers were estimating the hours required to complete each job. He referred to this process as an over/under-analysis and provided a few steps to start our work concerning exactly which variables he wanted us to analyze and which could be ignored. However, most of the data needed to be 'cleaned' before we could perform any analyses using XLMiner since there are certain restrictions for each type of test.

Therefore, our first goal was to clean and organize the data, such as removing any unnecessary variables, handling missing data, eliminating major outliers, etc. This also included going through variables, like the "service types," and narrowing them down from 460 different types to the most common nineteen by using a pivot table to show number of occurrences (Appendix A). Additionally, the same system was applied to the "crew/assigned to" variable, which we decreased from 164 different crews/managers and combinations to the fifteen most frequently used (Appendix B). Next, since our primary goal was to look at the hours budgeted versus the hours it actually took to complete the job, we created a new column of budgeted minus recorded hours to show the difference between the two. We then created a histogram using MegaStat in Excel in order to visualize how the data were distributed and to determine the range of hours on which we should focus (Appendix C). This allowed us to ensure that the data were normally distributed, and then we were able to bin the budgeted minus recorded hours into categories falling between -10 and +10. However, this would only give Andrew a very broad picture of how effectively they were budgeting their hours for each job site. Therefore, we created a tight-binned column with hours categorized from -2 to +2 in order to give him a more accurate distribution. These parameters are specifically entailed in Appendix E.1, and on the "Testing Future Jobs" worksheet.

Finally, we were able to run Naive Bayes Classifier (NBC) on the cleaned and partitioned data, allowing us to classify any new jobs by finding similar jobs. Essentially, the tables created using NBC represent the probability that each category (e.g., March, 79556, etc.) for any attribute (e.g., month, zip code, etc.) fall into each class. Furthermore, these probabilities were used to create a "Testing Future Jobs" formula that calculates an overall prediction that a job would be under- or

over-budgeted according to the job's attributes. After presenting the data analysis and predictions to Andrew, he appreciated our work but requested that we make an even tighter binned category of -3 to +3 in order to focus more clearly on the hours previously classified as "0." The parameters for this second NBC test are found in Appendix E.2, and on the "Testing Future Jobs (2)" worksheet.

Results & Discussion

Based on our data analysis, we found that MasterScapes is very good at estimating job completion times. According to Naive Bayes Classifier, 97% of jobs are completed within the budgeted time. Looking at performance according to service, MasterScapes tends to under-budget mowing jobs most often. This is likely due to the fact that the majority of services performed by MasterScapes are mowing jobs. However, right after mowing, we found that ornamental care, bed cleaning, trimming, and irrigation repairs are the services most likely to be under-predicted. We recommend that MasterScapes focus on improving efficiency for mowing first, because mowing is the most popular service. After mowing efficiency has improved, the four additional services (ornamental care, bed cleaning, trimming, and irrigation) should be addressed.

As far as work crews, the APA team consistently performs jobs faster than any other crew. 71% of their jobs are classified as being finished faster than the budgeted time. Based on our results, the Sweetwater team is the slowest performing team at MasterScapes with almost 50% of jobs being under-budgeted. It is important to consider that SW1 is the only Sweetwater team, so the other 50% of their jobs are classified as being within the budgeted hours. Even though half of Sweetwater's jobs are within the budgeted hours, they are still significantly behind every other

MasterScapes work crew. It is especially important to improve Sweetwater's accuracy because they are the only work crew in their area. Right behind SW1 is crew 1006 as the slowest performing crew in Abilene. 25% of 1006's jobs are classified as being under-budgeted. This is significant because the other 6 mowing crews are performing at less than 12% under-budgeted jobs. Because the other mowing crews are performing the same services at a much faster rate, this data could indicate an issue within the 1006 team.

Next, we analyzed the results according to month. May is the busiest Month for MasterScapes, this is not surprising because Spring is the best season for landscaping. However, this means that May has the most under-budgeted out of any other month. March and April are right behind May as having the most percentage of under-budgeted jobs. Again, this is due to the season. Our results conclude that January is the most accurate month for Masterscapes, meaning it is the month that contains jobs most consistently classified as 0. Finally, October has the highest probability of being over-budgeted, making it the most efficient month in our data. 75% of jobs completed in October fall into the 1 category, which means that they are finished 3-10 hours faster than what was originally budgeted. 28% of October jobs fall into the 2 category, which means that they were completed 10 or more hours faster than the estimated time. We recommend that MasterScapes further adjust the budgeted hours for March-May. This is the busiest time of the year, which means it will take longer to complete jobs. Therefore, MasterScapes should increase the average estimated time during these months by several hours in order to improve overall accuracy.

Finally, we were asked to analyze MasterScapes' performance based on location. 79556 appears to be the busiest zip code, meaning most of MasterScapes' clients live in this area. This also means that 79556 has a large percentage of under budgeted jobs. 79606 is the next slowest zip

code with nearly 53% of jobs likely to be classified as being under budgeted. On the other hand, 79602 is the more efficient than most zip codes with 93% of jobs classified as either completed on time or faster than predicted. Only 15% of jobs in 79602 are classified as being under budgeted. After looking at zip code, we were asked to analyze performance by address. Based on our results, we found that 400 E Broadway St and 104 Southview Terrace have the most under budgeted jobs. 400 E Broadway St (Texas National Bank, Sweetwater, TX) is the most frequent with 29 jobs classified as finished late, and 104 Southview Terrace (private residence, Sweetwater, TX) is the second most with 23 jobs under budgeted. The third most occurring is 128 Southview Terrace (private residence, Sweetwater, TX), with only 13 jobs finished late. This indicates that Texas National Bank and the 104 Southview Terrace home are significantly higher than any other address. It is important to note that all three of the most frequently under budgeted addresses in the MasterScapes database are all in Sweetwater. This is further evidence that the Sweetwater branch needs to increase budgeting accuracy.

Bibliography

Taylor, Bernard W., and Laurence J. Moore. "Estimating Job Flow-times in a Job Shop for Contractually Negotiated Due-dates". *The Journal of the Operational Research Society* 33.9 (1982): 845–854. Web.

Appendices

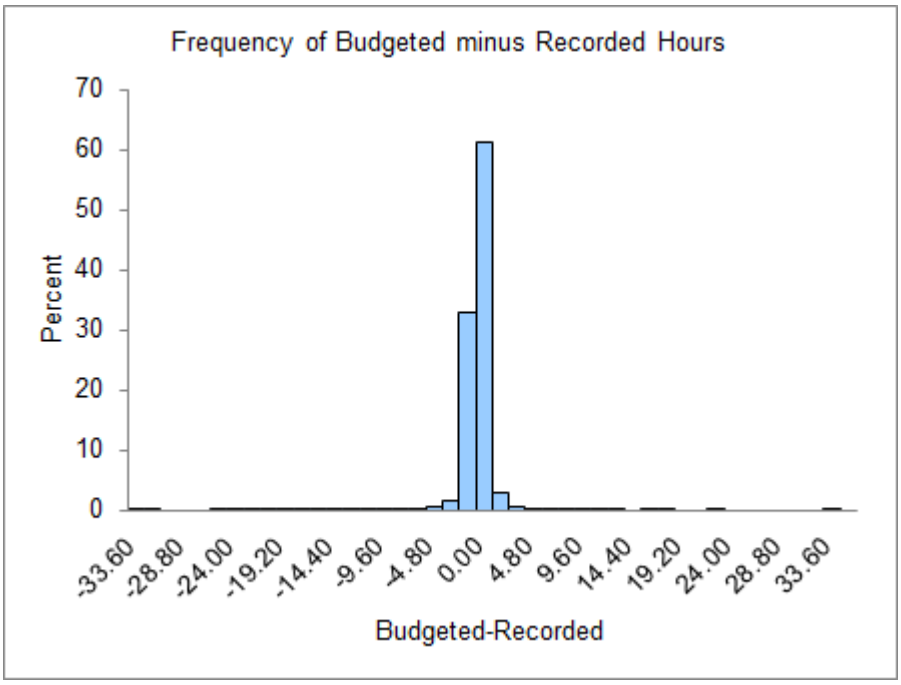
Appendix A

ID number	Crew/Assigned To:	Frequency
1	1001	1862
2	1002	1797
3	1006	1718
4	1005	1706
5	SW1	1665
6	1007	1593
7	APC	1506
8	APB	1491
9	OR2	1484
10	1003	1372
11	APA	1349
12	OR1	1105
13	APD	1011
14	1004	998
15	IR1	711

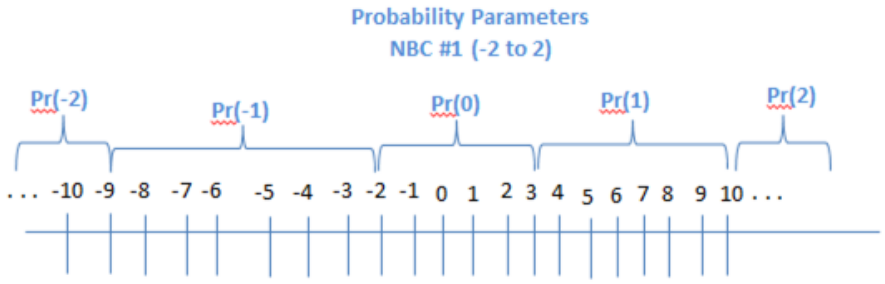
Appendix B

Address	# of under budgeted jobs
400 E Broadway St	29
104 Southview Terrace	23
128 Southview Terrace	13
310 FM 1856	12
300 SE Georgia Ave	9
3525 Rolling Green Dr	8
5410 Buffalo Gap Rd	8
6543 E Hwy 180	8
Hwy 277 S and CR 403	8
401 E Broadway St	7
302 SE Georgia Ave	6
1209 Hwy 70	5
1700 E 12th St	5
38 Muirfield St	5
409 CR 119	5
733 EN 16th St	5

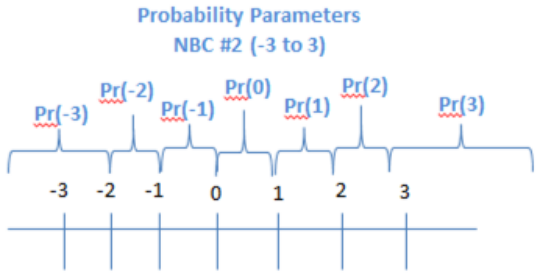
Appendix C



Appendix E.1



Appendix E.2



Appendix F

CREW	-2	-1	0	1	2
1001	0	0.096153846	0.090983859	0.043010753	0
1002	0	0.025641026	0.08944176	0.086021505	0
1003	0.0625	0.070512821	0.061169939	0.043010753	0
1004	0	0	0.052225763	0.032258065	0
1005	0	0.019230769	0.087488434	0.043010753	0
1006	0.25	0.064102564	0.080394777	0.010752688	0
1007	0.125	0.057692308	0.081114424	0.075268817	0
APA	0	0	0.055823995	0.021505376	0.714285713
APB	0	0	0.061786779	0.032258065	0
APC	0	0	0.064768171	0.010752688	0
APD	0	0.006410256	0.040608615	0.043010753	0
IR1	0.0625	0.044871795	0.026935335	0.053763441	0
IR2	0.0625	0.044871795	0.013159247	0.032258065	0
OR1	0	0	0.054179089	0.021505376	0
OR2	0.0625	0.025641026	0.07515164	0.129032258	0
SW1	0.375	0.544871795	0.064768171	0.322580645	0.285714285

Appendix G

SERVICE	-2	-1	0	1	2
1	0.4375	0.397435897	0.451835098	0.430107527	0.285714285
2	0	0	0.147219081	0.075268817	0.142857143
3	0.0625	0.141025641	0.104451527	0.129032258	0
4	0.0625	0.141025641	0.083170556	0.010752688	0
5	0	0.006410256	0.065590624	0.032258065	0.571428571
6	0.125	0.051282051	0.039169323	0.032258065	0
7	0.125	0.121794872	0.02919708	0.096774194	0
8	0	0	0.004009458	0	0
9	0.125	0.064102564	0.015112573	0.043010753	0
10	0	0	0.017991159	0.075268817	0
12	0	0	0.0155238	0	0
14	0.0625	0.051282051	0.008841369	0	0
15	0	0	0.002467359	0.032258065	0
16	0	0	0.005962784	0	0
17	0	0.019230769	0.004934718	0.043010753	0
19	0	0.006410256	0.004523491	0	0

Appendix H

MONTH	-2	-1	0	1	2
1	0	0.006410256	0.031767246	0.075268817	0
2	0	0.006410256	0.037935643	0.086021505	0.285714285
3	0.25	0.141025641	0.077927419	0.11827957	0
4	0.125	0.314102564	0.130358795	0.075268817	0
5	0.4375	0.192307692	0.129022309	0.11827957	0.285714285
6	0.125	0.179487179	0.133751414	0.11827957	0
7	0.0625	0.08974359	0.145265755	0.086021505	0.142857143
8	0	0.025641026	0.150920119	0.107526882	0
9	0	0.032051282	0.118741647	0.139784946	0
10	0	0.012820513	0.044309654	0.075268817	0.285714285

Appendix I

ZIP	-2	-1	0	1	2
75920	0	0	0.000411226	0	0
76424	0	0	0.000719646	0.010752688	0
76437	0	0	0.002056132	0	0
76443	0	0	0.000411226	0	0
76448	0	0	0.00061684	0	0
76834	0	0	0.00030842	0	0
76933	0	0.038461538	0.002570166	0.032258065	0
79504	0	0.006410256	0.005243138	0	0
79506	0	0.012820513	0.001130873	0	0
79508	0	0.006410256	0.001953326	0.010752688	0
79510	0	0	0.00092526	0	0
79520	0	0	0.000102807	0	0
79525	0	0	0.000719646	0	0
79526	0	0.012820513	0.002261746	0	0
79530	0	0	0.00030842	0	0
79532	0	0	0.000102807	0	0
79536	0	0	0.001233679	0.010752688	0
79541	0	0	0.004934718	0	0
79543	0	0.006410256	0.000411226	0	0
79545	0	0.006410256	0.002158939	0	0
79546	0	0	0.001644906	0	0
79549	0.0625	0.064102564	0.00853295	0	0
79556	0.3125	0.429487179	0.064768171	0.344086021	0.428571428
79562	0	0	0.033412152	0	0
79563	0	0	0.00092526	0	0
79567	0	0	0.00030842	0	0
79601	0	0.102564103	0.170350571	0.107526882	0
79602	0.0625	0.08974359	0.187519276	0.182795699	0.571428571
79603	0	0	0.022720263	0.043010753	0
79605	0.1875	0.064102564	0.192762414	0.064516129	0
79606	0.375	0.16025641	0.285699599	0.161290323	0
79697	0	0	0.002775779	0.032258065	0