The County Business Advisor

Sandeep Kumar Moparthy, Pooja Karla, Janit Modi, Jaebin Park

Data Preprocessing

As mentioned in the project proposal, the County Business Partners included eight different datasets. For the purposes of our analysis, we decided to focus on three: the Complete County, State, and US files. Each dataset included around 20 columns, many of which were either irrelevant or unusable. For example, many features, such as Annual Payroll for the Industry, had supporting variables, like Annual Payroll Noise Flag, that was formatted in way that made it difficult to include in the analysis. Thus, we moved forward by removing these columns.

The next set of variables that needed managing were the Number of Establishments in each Industry by Employee Size Class. As it was dealing with size, there were at least a dozen columns, each representing a different size class, from 1-5 to 100-499 to 5000+. Because one of our objectives for the analysis is to provide companies with insight on not only where some industries thrive, but also what size companies are typically present in those areas, we decided to create three classes: small, medium, and large. Establishments with 1-99 employees were included in the small class, those with 100-999 employees in the medium class, and the rest in the large class. With this data, it could be possible to receive the output that companies want through classification.

Finally, in order to create a central feature that we can use to combine our multiple datasets, we needed to clean the NAICS industry codes. The issue was that these codes were given as factors with either multiple hyphens (“-“) or back slashes (“/”). To make this variable usable, we converted each index into a string, replaced the characters that were not numbers into an empty space, and converted it back into an integer. We ran into a problem when performing this function. Our dataset was too large (the County dataset has over two million rows) and it was computationally taking too long. We decided to randomly sample our dataset to *1/sqrt(n)* to make it much more manageable.

We performed these steps for the three datasets. Because some datasets have features that others don’t, some data specific preprocessing needed to be done. After which, we combined the State and County files by NAICS and the FIPS State Code, with which we used the NAICS to merge the US file to create a master set that will be using for further modeling and analysis. Once we had our master dataset, some variables were no longer necessary, leading us to drop the features NACIS, State_lfo, and US_lfo. We concluded processing the data by creating separate list that contained the Average Annual Payroll for the County, State, and US. Our final dataset includes the following features:

COUNTY_emp: County total mid-march employees

COUNTY_qp1:	County first quarter payroll
COUNTY_ap:	County annual payroll
COUNTY_est:	County total number of establishments
COUNTY_sz_small:	County establishment class size, small
COUNTY_sz_med:	County establishment class size, medium
COUNTY_sz_lrg:	County establishment class size, large
state_emp:	State total mid-march employees
state_qp1:	State first quarter payroll
state_ap:	State annual payroll
state_est:	State total number of establishments
state_sz_small:	State establishment class size, small
state_sz_med:	State establishment class size, medium
state_sz_lrg:	State establishment class size, large
US_emp:	US total mid-march employees
US_qp1:	US first quarter payroll
US_ap:	US annual payroll
US_est:	US total number of establishments
US_sz_small:	US establishment class size, small
US_sz_med:	UA establishment class size, medium
US_sz_lrg:	US establishment class size, large

Modeling

We decided to create an initial model with just the variables regarding Counties from the state Texas. Our reasoning behind this logic was that Counties is the most specific compared to State and US. Furthermore, Texas was the state that contained the most counties, allowing us to test on a smaller sample and still perform a relatively accurate analysis. For our final analysis, we intend to fully automate the process, where the client only has to provide a few factors, such an intended average payroll, the establishment capacity, or the industry type that they're involved in, and our model will run through the entire data set and provide us with all the necessary information.

As it is our first model, we fitted a Multiple Linear Regression on the Average Payroll for the Counties on all the variables from the County dataset. Naturally, we split our dataset into training and test sets. We fit the model with the training set, predicted using the test set. We calculated the mean squared error, but it outputted a very high value, telling us that we need to revise our model.

For further modelling and analysis, we plan on keeping our response variable, the average payroll, the same. As mentioned before, we intend on receiving some information from our clients, with which we will run our model. The output will contain the mean and median for the avg. payroll for either the county, state, or us for the specified industry. To obtain this output, we currently plan on incorporating some clustering or classification methods, such as hierarchical clustering or random forest, to classify by location. A combination of the location with the highest avg. payroll, the

selected industry, and ideal establishment capacity will be output that we will provide to our clients.

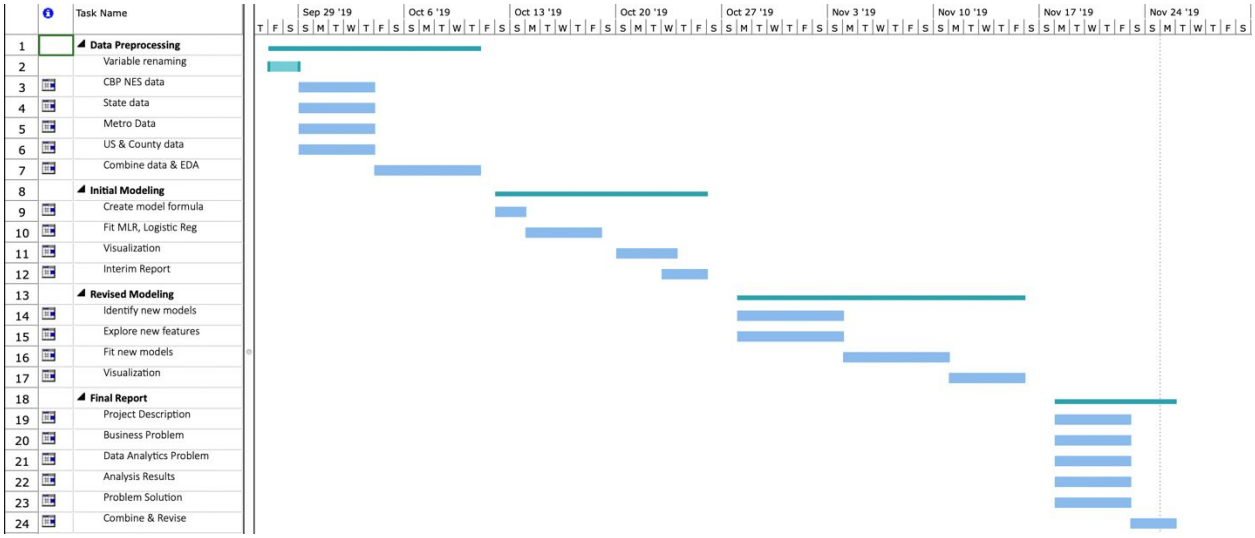
The reason why we've spent much time on the data preprocessing and initial modelling steps is because we can treat our data in a hierarchical fashion, where changing the model will only take small changes in scale. For example, if a client wishes to know about the agriculture industry for a specific state and establishment size, we can shift from the county level to the state level and choose from the three size categories that we've created.

Gantt Chart & Changes

As of right now, our project is on track with the original Gantt chart from the project proposal. The only thing that is missing is the initial model visualization, which we believed to be not necessary at this stage, since the graphs will be rough and not contain much valuable information.

We're slightly ahead of schedule, as we've already begun exploring new models and improvements that we can make on our initial model. Following is the updated Gantt chart with the start/finish dates as well as the completion percentage of the tasks.

	Task Name	Duration	Start	Finish	Contact	% Complete
1	**Data Preprocessing**					**100%**
2	Variable renaming	2 days	Fri 9/27/19 8:00 AM	Sat 9/28/19 5:00 PM	Jaebin	100%
3	CBP NES data	5 days	Sun 9/29/19 8:00 AM	Thu 10/3/19 5:00 PM	Sandeep	100%
4	State data	5 days	Sun 9/29/19 8:00 AM	Thu 10/3/19 5:00 PM	Janit	100%
5	Metro Data	5 days	Sun 9/29/19 8:00 AM	Thu 10/3/19 5:00 PM	Pooja	100%
6	US & County data	5 days	Sun 9/29/19 8:00 AM	Thu 10/3/19 5:00 PM	Jaebin	100%
7	Combine data & EDA	7 days	Fri 10/4/19 8:00 AM	Thu 10/10/19 5:00 PM	Sandeep	100%
8	**Initial Modeling**					**100%**
9	Create model formula	2 days	Sat 10/12/19 8:00 AM	Sun 10/13/19 5:00 PM	Sandeep, Jaebin	100%
10	Fit MLR	5 days	Mon 10/14/19 8:00 AM	Fri 10/18/19 5:00 PM	Sandeep	100%
11	Interim Report	3 days	Wed 10/23/19 8:00 AM	Fri 10/25/19 5:00 PM	Jaebin	100%
12	**Revised Modeling**					**13%**
13	Identify new models	7 days	Mon 10/28/19 8:00 AM	Sun 11/3/19 5:00 PM	Sandeep, Jaebin	50%
14	Explore new features	7 days	Mon 10/28/19 8:00 AM	Sun 11/3/19 5:00 PM	Janit, Pooja	0%
15	Fit new models	7 days	Mon 11/4/19 8:00 AM	Sun 11/10/19 5:00 PM	Everyone	0%
16	Visualization	5 days	Mon 11/11/19 8:00 AM	Fri 11/15/19 5:00 PM	Everyone	0%
17	**Final Report**					**0%**
18	Project Description	5 days	Mon 11/18/19 8:00 AM	Fri 11/22/19 5:00 PM	Pooja	0%
19	Business Problem	5 days	Mon 11/18/19 8:00 AM	Fri 11/22/19 5:00 PM	Pooja, Janit	0%
20	Data Analytics Problem	5 days	Mon 11/18/19 8:00 AM	Fri 11/22/19 5:00 PM	Sandeep, Jaebin	0%
21	Analysis Results	5 days	Mon 11/18/19 8:00 AM	Fri 11/22/19 5:00 PM	Sandeep, Jaebin	0%
22	Problem Solution	5 days	Mon 11/18/19 8:00 AM	Fri 11/22/19 5:00 PM	Janit	0%
23	Combine & Revise	3 days	Sat 11/23/19 8:00 AM	Mon 11/25/19 5:00 PM	Everyone	0%


Task Name
Sep 29 '19
Oct 6 '19
Oct 13 '19
Oct 20 '19
Oct 27 '19
Nov 3 '19
Nov 10 '19
Nov 17 '19
Nov 24 '19
1 Data Preprocessing
2 Variable renaming
3 CBP NES data
4 State data
5 Metro Data
6 US & County data
7 Combine data & EDA
8 Initial Modeling
9 Create model formula
10 Fit MLR, Logistic Reg
11 Visualization
12 Interim Report
13 Revised Modeling
14 Identify new models
15 Explore new features
16 Fit new models
17 Visualization
18 Final Report
19 Project Description
20 Business Problem
21 Data Analytics Problem
22 Analysis Results
23 Problem Solution
24 Combine & Revise