Analytics Capstone

Document Analysis (Evaluation and Deployment)

Predictability of Career and Professional Excellence

Group 4

Introduction:

In the Evaluation and Deployment phase, we will assess the results obtained from our data analytics endeavors and devise a strategy for deploying the insights gained from the project. By evaluating the performance of our models against business objectives and reviewing the overall process, we aim to determine the next steps for maximizing the value derived from this dataset. Each posting includes 27 valuable attributes, ranging from job titles and descriptions to salaries, locations, and application URLs. Additionally, separate files contain information on benefits, skills, industries, and associated companies, providing a rich source for exploration and analysis.
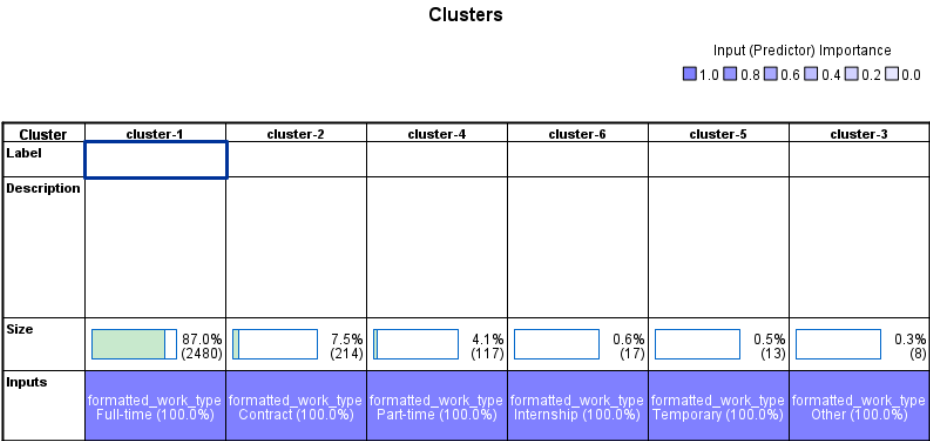
A. Evaluation:
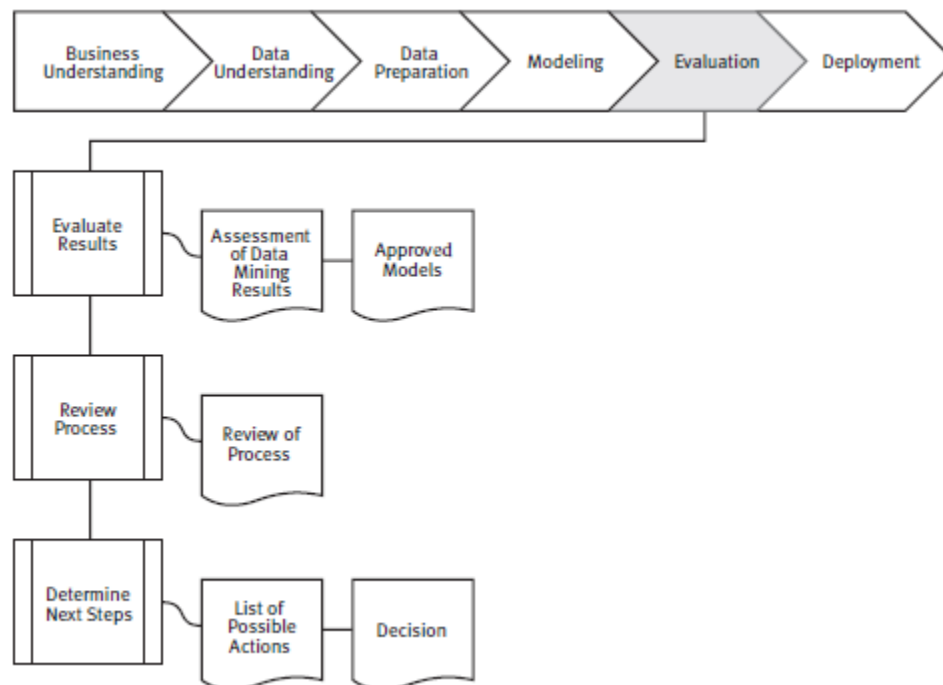
1. Evaluate results:
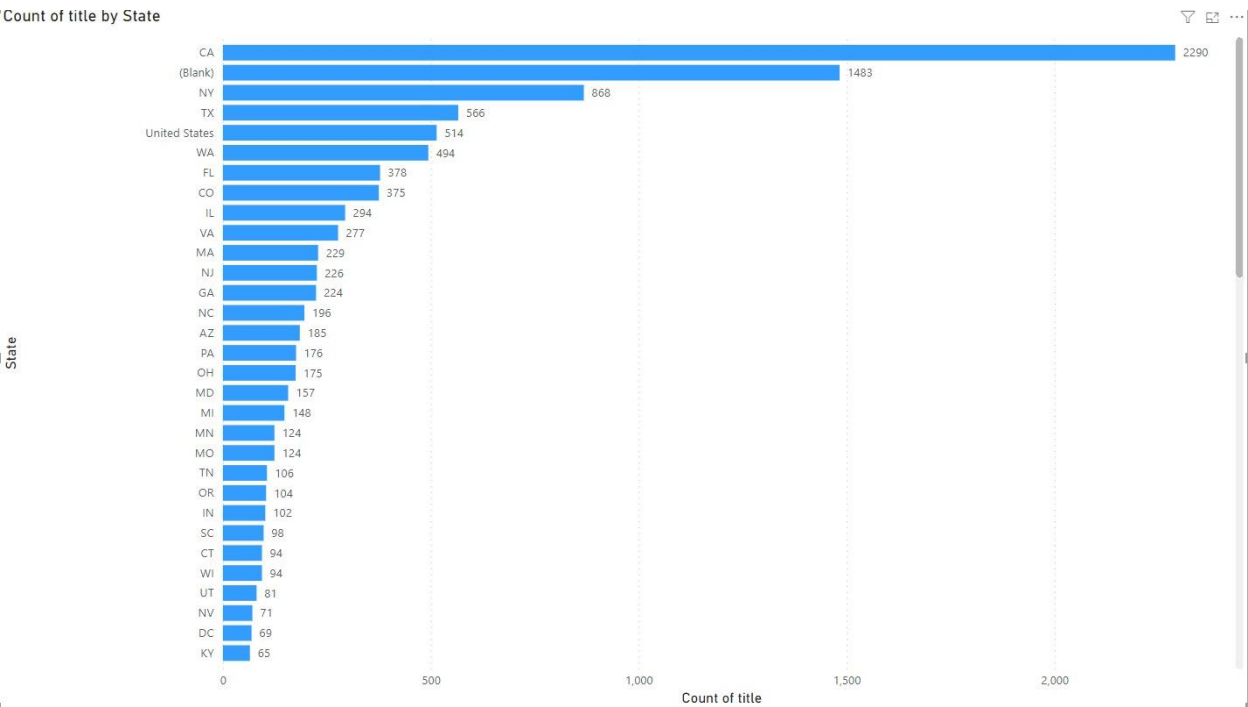
Process: During this phase, we will assess the performance of our models in meeting the business objectives outlined for the project. This evaluation will focus on factors such as accuracy, generality, and alignment with business success criteria. Additionally, we will examine other data mining results generated to identify potential challenges, information gaps, or future directions.

Model 1: K-means Clustering

We assessed the initial model, which relies on the variables "*formatted_work_types*" and "formatted_experience_level." Using this data, we segment it into six clusters.

Clusters

Input (Predictor) Importance
□1.0 □0.8 □0.6 □0.4 □0.2 □0.0

Cluster	cluster-1	cluster-2	cluster-4	cluster-6	cluster-5	cluster-3
Label						
Description						
Size	87.0% (2480)	7.5% (214)	4.1% (117)	0.6% (17)	0.5% (13)	0.3% (8)
Inputs	formatted_work_type Full-time (100.0%)	formatted_work_type Contract (100.0%)	formatted_work_type Part-time (100.0%)	formatted_work_type Internship (100.0%)	formatted_work_type Temporary (100.0%)	formatted_work_type Other (100.0%)

Output: We decided to employ this model to delineate the spectrum of work types with respect to varying levels of experience.

'Count of title by State

State	Count of title
CA	2290
(Blank)	1483
NY	868
TX	566
United States	514
WA	494
FL	378
CO	375
IL	294
VA	277
MA	229
NJ	226
GA	224
NC	196
AZ	185
PA	176
OH	175
MD	157
MI	148
MN	124
MO	124
TN	106
OR	104
IN	102
SC	98
CT	94
WI	94
UT	81
NV	71
DC	69
KY	65

The team also evaluated job posting frequency across various states, providing valuable insights into the geographical focus of companies in attracting applicants. It

was observed that most job postings are concentrated in populous states, which aligns with the distribution of potential applicants.

Model 2: Regression for the prediction of average salary

While evaluating the model we found that there was no error and while analyzing more in detail found that the model failed as it never gave an error or no less than 99% prediction.

Comparing $E-average _salary with average _salary

Minimum Error	-0.0
Maximum Error	0.0
Mean Error	-0.0
Mean Absolute Error	0.0
Standard Deviation	0.0
Linear Correlation	1.0

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	min_salary, max_salary[b]	.	Enter

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	1.000	1.000	.002831

a. Predictors: (Constant), min_salary, max_salary

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.730E+13	2	2.865E+13	.	.[b]
	Residual	.089	11108	.000		
	Total	5.730E+13	11110			

b. Predictors: (Constant), min_salary, max_salary

Coefficients

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-2.910E-11	.000		.000	1.000
	max_salary	.500	.000	.610	624193565.54	<.001
	min_salary	.500	.000	.409	418284118.44	<.001

Output: We decided *not to* use this model again as it is a failure.

Model 3: Linear Modelling.

After experimenting with various iterations of the model, it was determined that the second version exhibited superior predictive accuracy compared to the other two versions. This conclusion was drawn following an evaluation of the model's performance.

| Version 1.0 | Version 2.0 | Version 3.0 |
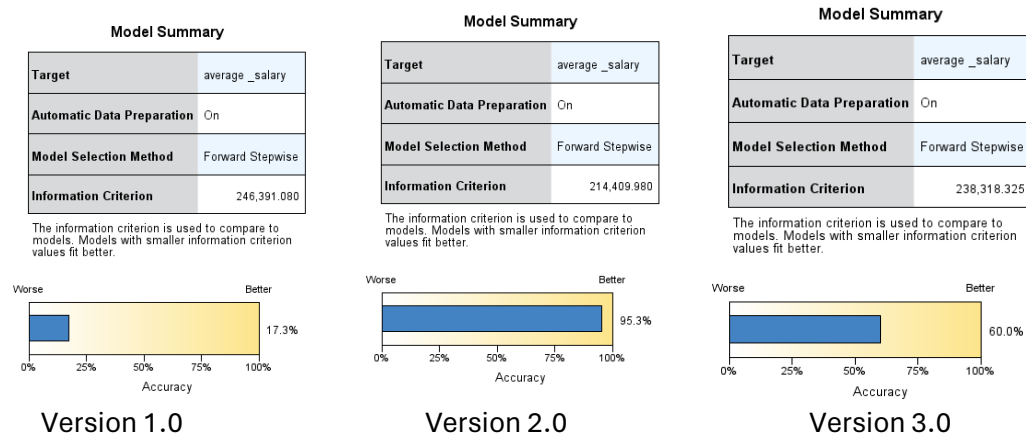
As we can clearly see in the model summary 2.0 has the best accuracy among all three model to predict the average salary, but we also know that model 2 has more inputs that are related to the average salary the max and min salary inputs thus it has more accuracy

Output: Thus, we are going to use the model 2.0 & 3.0 and test it to know which is better.

Model 4: **CHAID** Model to predict whether the job offered is remote job or not.
While analyzing this model, we concluded it is good to predict the job to get allowed for remote or no.
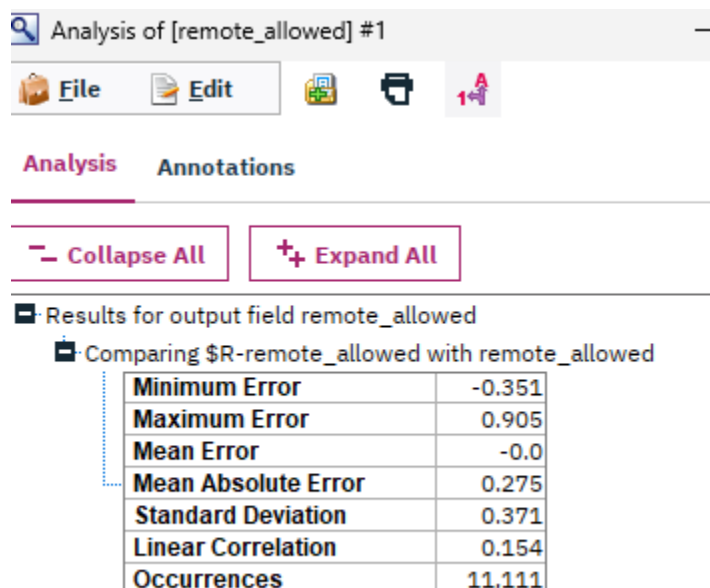
Output: We will summarize the assessment results in terms of business success criteria, determining whether the project has successfully met its initial objectives. Approved models that meet the selected criteria will be identified for further consideration.

2. Review process:

Process: A thorough review of the data mining engagement will be conducted to identify any overlooked factors or tasks. This review will encompass quality assurance issues to ensure that models were built correctly and that only permissible attributes were utilized.

While reviewing the data we cleaned it up more thoroughly that we came to a final data of 11112 records, and we decided to use it to create different models and then deploy them to make the final dashboard.

Output: We will summarize the process review, highlighting any missed activities and identifying tasks that may need to be repeated for accuracy and completeness.

3. Determine the next steps:
Process: Based on the assessment and process review outcomes, the project team will decide on the appropriate course of action. This decision may involve concluding the project and proceeding with deployment, initiating further iterations for refinement, or initiating new data mining projects. Considerations such as remaining resources and budget constraints will inform these decisions.

Output: We will list potential further actions along with the reasons for and against each option, facilitating an informed decision-making process. The decision regarding how to proceed will be described, accompanied by a rationale that considers the assessment results, process review findings, and project objectives.

B. Deployment
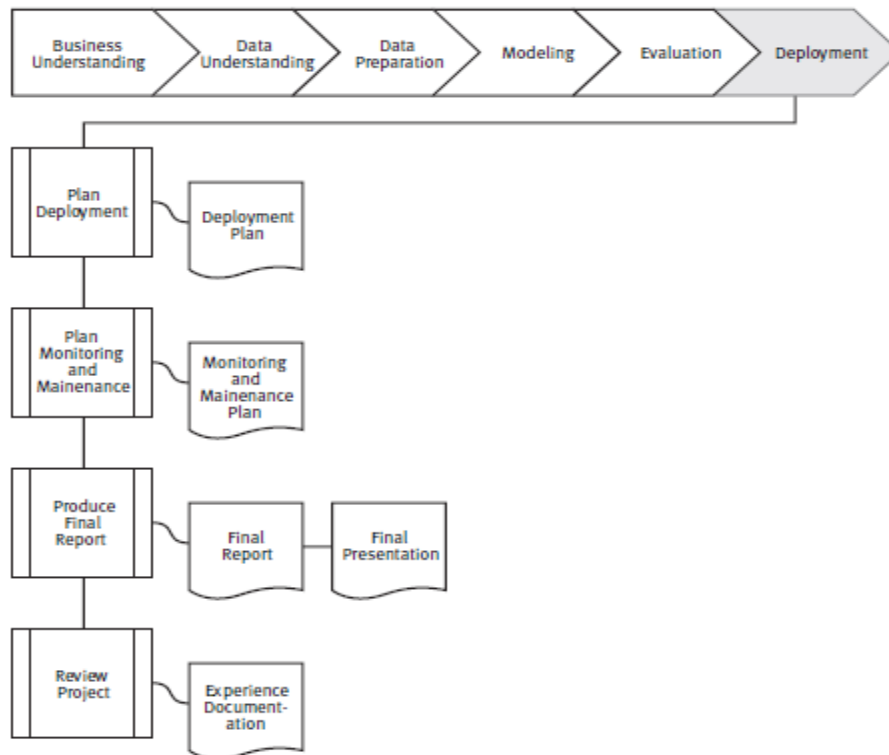
1. Plan Deployment

Process: Drawing from the evaluation results, we will devise a strategy for deploying the insights gained from the project. This plan will outline the necessary steps for implementing the relevant models and leveraging the insights to drive business value.

- **Output:** The deployment plan will summarize the strategy, detailing the necessary steps and procedures for executing the deployment process effectively.

    Model Selection for Deployment:
    - o Based on the evaluation results, it was determined that Model 2.0, utilizing regression for predicting average salary, demonstrated

superior predictive accuracy compared to other models. Additionally, Model 3.0, employing linear modeling, also showed promising results.

o Therefore, the recommended approach is to deploy Models 2.0 and 3.0 for further testing and refinement.

o These insights can be used by the applicants to modify their job-search based on the job-type and the salary range for those jobs.

o It will also help them implement these results in their decision-making process to help them make more data-driven decisions and focus on the desired goals.

Conclusion

In conclusion, the Evaluation and Deployment phase plays a crucial role in ensuring that the insights derived from the LinkedIn job postings project are effectively utilized to drive business outcomes. By evaluating the results, reviewing the process, and determining the next steps, we can maximize the value derived from the project and pave the way for successful deployment and implementation of the insights gained.