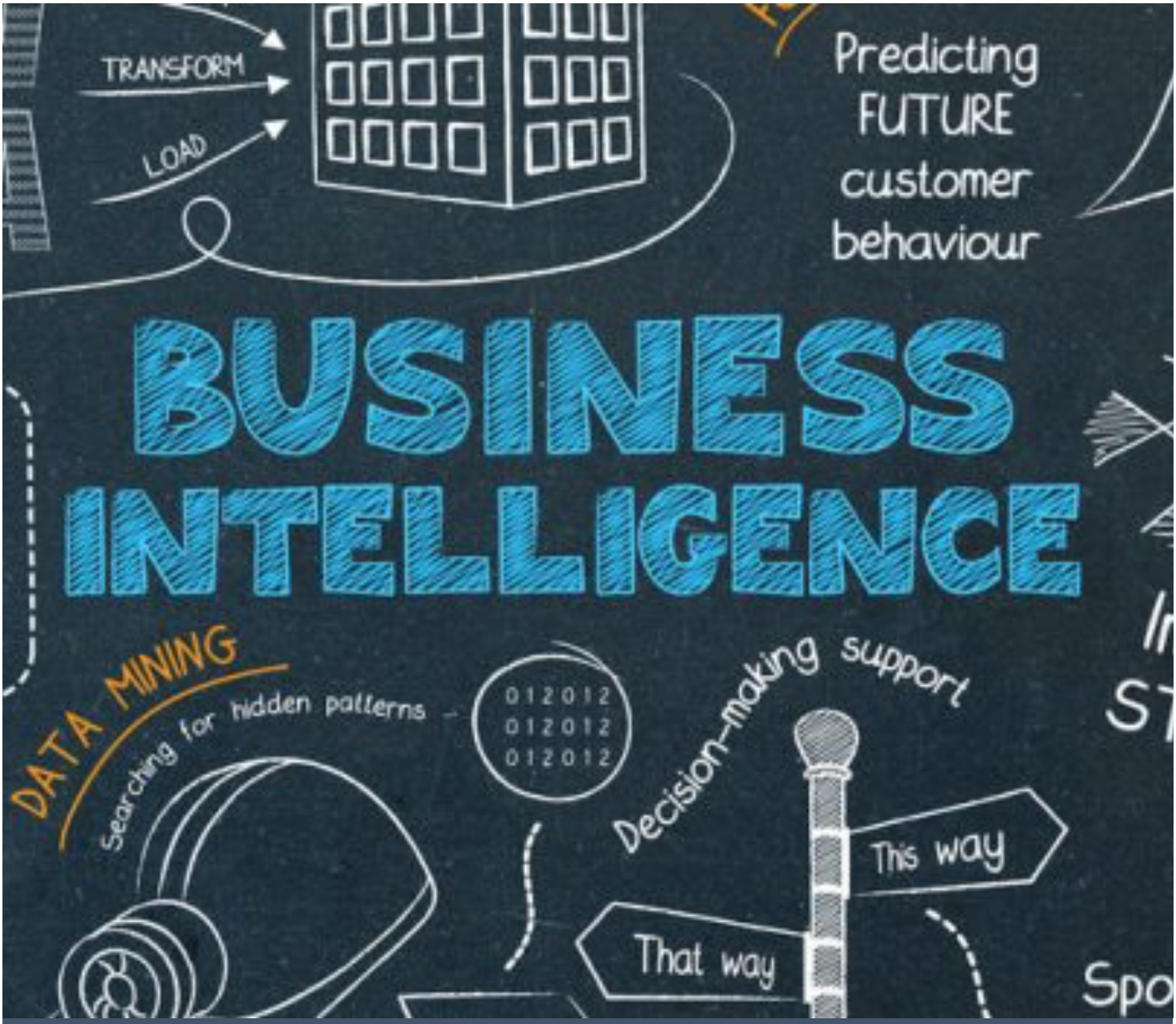


Deployment, Monitoring and Maintenance

Project Milestone 7

Dithapelo Huma. Maishibe Thobela

5/13/21

BIN381

Table of Contents

Table of Figures

Plan deployment

Introduction

The deployment phase is the stage in the process where we use what we discovered to improve the business. In order to successfully deploy a model, one must understand the requirement. The objective of this project was to the use demographic data of employees, in order to build a model that can accurately identify individuals whose salary exceeds R50000. This was so that the sponsor can save money on his marketing efforts, so that they can offer their service to this kind of individuals only, assuming that this is an individual that is likely to buy

Deployable results.

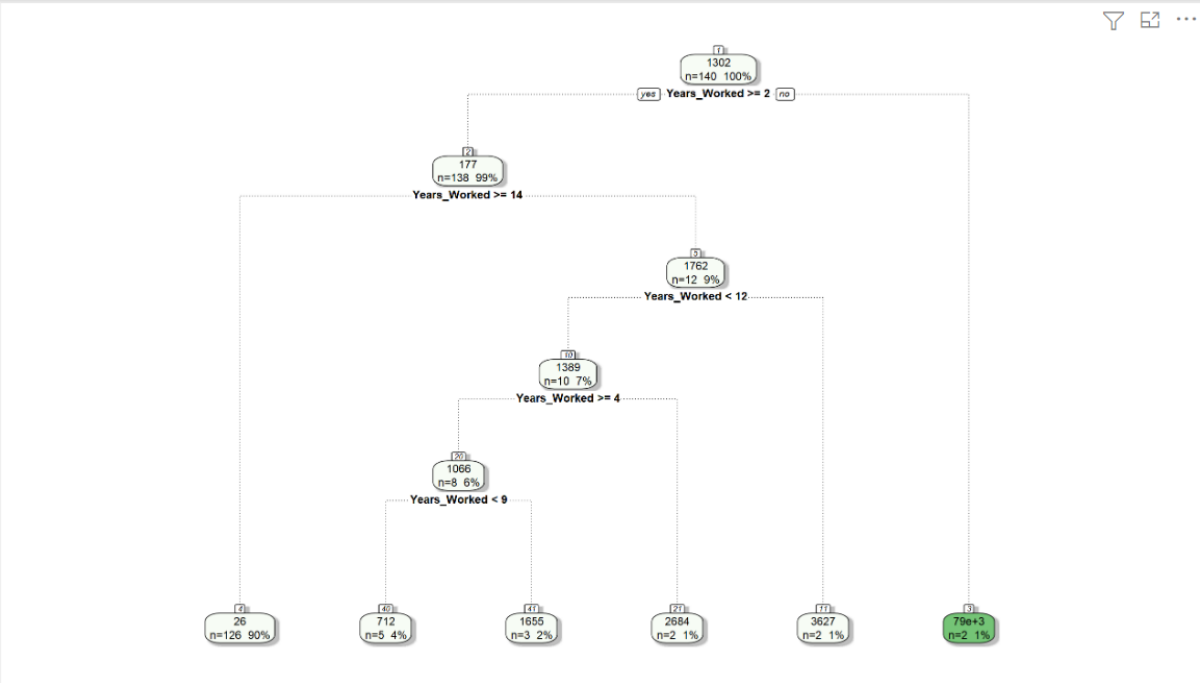
The deployable results are the results that we used to build our final models. We used a neural network model as well as a tree diagram model. The tree diagram model produced more accurate results at 95%, and it will therefore be the model that we deploy, as we want to give the sponsor the most accurate model.

According to our tree diagram model, the attributes that produce the most accurate model of individuals who earn more than R50000, were a combination of the categorical/ nominal variables we created and variables that were given. They include:
Department name (which we changed to a factor data type to improve accuracy)

- Above50K
- SalaryType
- AgeGroup
- Years_Worked

DepRarity (which was an attribute we added to improve accuracy)

- This model compared employees who earned above R50k with the amount of years they worked. It was able to march 54042 employees to their salary and the length of their employment.
- The accuracy of the model is at 0.9662 (96.62%)and the confidence interval at 95% is [0.9647, 0.9677]
- The kappa value of this model stands at 0.507, which show a very good agreement between internal processes/observers. This means that this model is more reliable than the first model based on the difference in the kappa values.
- However, the P-Value is at 2.2e-16. this shows a very significant correlation between employees who earn above R50k and the length of their employment.
- Based on the kappa value and the P-value, this model provides is with information that we can rely on. The P-value suggests that there is a significant correlation between the number of years that employees have worked and whether or not they earn above R50 000. The kappa value suggests that should an independent entity review the second model, then we would be able to agree on both our results.



Alternative plans for deployment

We build two models for tree diagrams (see previous milestones) and a neural network model. The second most accurate model from the above is our other tree diagram model. In this model:

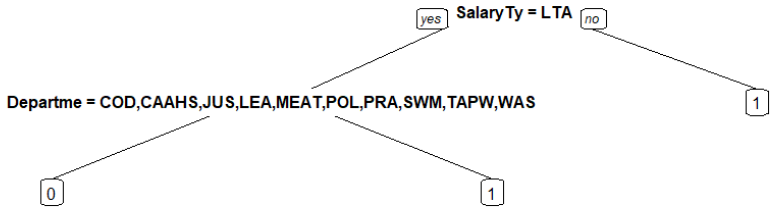
- The model was accurate in predicting that 54707 employees earned above R50k and those who worked at certain departments. It was however inaccurate in predicting the salary and department of 2303 employees
- The accuracy of the confusion matrix stands at 95%, and at a confidence interval of 95%, its accuracy lies between 0.9577 and 0.9609
- The Kappa value represents the agreement between two observers that are assessing the quality of their observations. With a Kappa value of 0, the internal comparison of the system did not agree with each other. If the Kappa value greater than 0.75, then we would conclude that the agreement would be excellent, and if it was below 0.4, it would show a poor agreement between observers
- With a P-value of 0.505, the is no statistical significance between the department that each employee works at and their salaries
- However, this model has a very low kappa value, meaning that two different entities would not be able to agree on the results they come up with, and due to the "Department.Name" being part of the attributes that were chosen for the first decision tree model, the p-value for this model was low. Showing us that the correlation between the employees that earned above R50 000 and the department that they worked for was very low
-

The variables we used for this model were:
Department name

- Above50K
- SalaryType
- AgeGroup

- Years_Worked

Although this model also had a high accuracy level, it was unreliable to use, because the kappa value was low. Kappa value is a value that tells us that if two people do the same thing the likelihood of producing the same results is low. A kappa value ensures reliability and consistency; thus, a low kappa value is not a good sign, hence this model was not our first deployment choice.



Distinct knowledge or information result

Employees that have worked less than 2 do not earn more than R50k
Veteran employees are not paid as well, as the new employees, which incentivizes the new employees to stay in the company.
There are specific departments where individuals that earn more than R50k work. R decided to abbreviate these departments when it sketched the tree diagram. There departments are:

- COD
- CAAHS
- JUS
- LEA
- MEAT
- POL
- PRA
- SWM
- TAPW
- WAS

From our results, we can conclude that there is a close correlation between the experience of the employees and the department they work in and whether or not they earn above R50k

What we learned from our first model; the tree listed the salary type “less than average” at the base of the tree. Employees who had a salary type of less than average would not be earning above R50 000. This made logical sense, as it would not make sense to have employees earning above R50 000 with a salary type of less than average. We can also see that employees who work under certain departments are more likely to earn R50 000. Departments that are not shown on the tree will not have employees who earn above R50 000. This can mean that certain departments might have a highly skilled workforce, which accounts for employees having to earn above R50 000

The first model was however disregarded as it had some issues. A second decision tree model which was based on the first model was created. The second model was more accurate than the first model. It allowed us to have a more detailed decision tree and we were able to gain valuable information from the model. It showed us the relationship between employees who earned R50 000 and how long those employees have worked at their respective jobs. Employees who worked less than 2 year would not be earning R50 000 or more. The longer and employee worked, the more likely they would be to earn R50 000 or more.

To an outsider, this shows that the company values employees who are loyal to it with high salaries. This can be pleasing to new employees as it would incentive working at a company for an extended period. This would add value to companies as they would be able to retain the workforce that they trust.

How knowledge will be propagated to users

We can build a site for the sponsors where they can have access to real live reports that we will generate from Power Bi and R. From this report, the sponsor will be able to continually have an updated list of individuals who earn more than R50K, as the employees' salaries increase due to performance, increase in skill, the number of years worked or even change departments

How the use of the result will be monitored and its benefits measured

The use of the results will be monitored by seeing how readily available the sponsors have access to an updated list of employees that earn more R50K, every time they need it. Our job as data analysts is to provide decision makers with the data that they need to make decisions that will save them money. Our ability to provide the sponsor with this information will help them make better decisions around the kind of individuals they will be targeting for their new service.

How the model result will be deployed within the organization systems

There different deployment strategies that exist, namely:

- Recreate
- Ramped
- Blue\Green
- A/B Testing
- Shadow

The deployment that we will think will be the best fit for this organization is "Blue\Green". This strategy is one where the new version is deployed alongside the older version. The strategy this organization used to use to determine the individuals that they can offer this new service is not mentioned- however it will be important to keep that running whilst developing the new version, for three main reasons:

- building the new version will take time
- it will allow us to have something to compare to, so that the sponsors can prove the value of the investment they made in data analysts
- it will allow for testing period

The switch to the new version/ way of doing things happens only when the new version is online and tested, and is safe for deployment. This is a safe option that comes with the benefit of having no downtime- which is important for any business. On the hand its disadvantage is how it is hard to implement and requires two running environments for a period, which will result in it being expensive. However, we think that the benefits of having a safe deployment and no downtime far outweigh the disadvantages, because it will allow management to have access to important information that will allow them to make decisions accurately, which will ultimately generate an income for the business.

How its use will be monitored and its benefits measured

We will be having different stages of monitoring. The initial monitoring will be done in our testing phase. This phase can be likened to a beta version. In this phase, we will start with having a limited number of users, that we can collect valuable feedback from and solve the problems that they experienced from using the site. This phase will allow us to monitor and measure benefits based on the experience of the initial users.

The second phase of monitoring our solution and measuring its benefit will be done overtime, by keep in contact with management, the financial and marketing department. From the marketing department, we will be able to monitor how our reports has helped them each right people, at the right time in a right way. From the financial department we will monitor the ROI that this investment has made. From management, we will receive continuous feedback, that will help us transform our data in more suited manner based on the organizations business needs, that will help them make informed decisions.

Plan Monitoring and Maintenance

With the completion and deployment of the model, monitoring and maintaining it will be a crucial activity as the performance of the model will have an influence on business practices and decisions. As data scientist we must be aware of the potential issues a model can face and we must be prepared to troubleshoot any issue that we may encounter.

Plan monitoring

Overtime models will degrade for various reasons, business policies may affect how long employees work for the company, departments may add or decrease the roles they offer employees and some departments may even be obsolete. We must therefore be knowledgeful of the types of data drifts that can occur and how we can overcome these drifts.

Model degradation

We have to monitor the following changes in the model:

Data Drift: A data drift occurs when new data/production data diverges from the training data that was used in the original model. A changing business environment can be the cause. For example. If we were to add a new department that hired young employees and paid them very high salaries, the model would start to produce less accurate results as the new data us not in uniform to the model.

Conceptual Data Drift: This occurs when the expectation of what constitutes a correct prediction. Overtime, the business will evolve and redefine its business practices. It will also change what it defines as a correct model prediction. For example, a company manufacturing microchips may start to disregard automakers as suitable clients due to a sudden and low consumer demand. Leaving this type of data drift unresolved, it may lead to us being unable to distinguish between two completely different attributes. (Laurent, 2020)

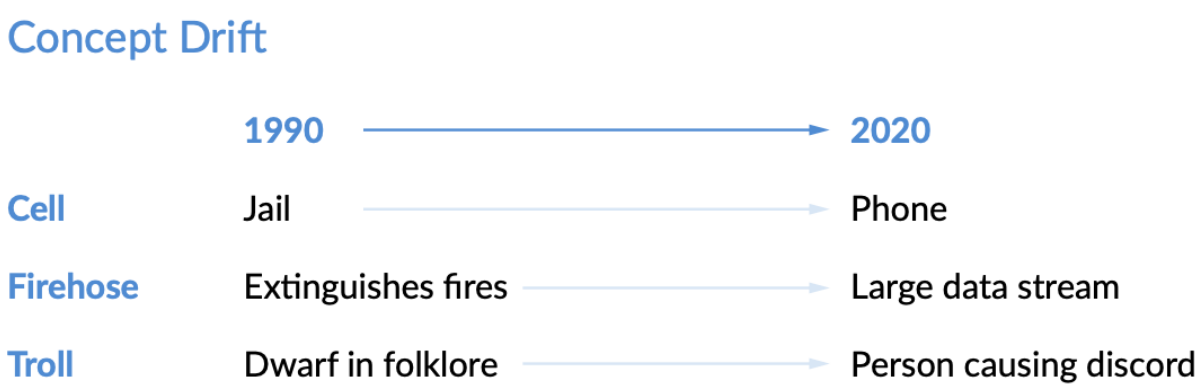


Figure 1: Concept drift

Figure 1 shows us how concept drift can lead to a complete misunderstanding of a model. Say for example a model was created in 1990 and the word "cell" was used to refer to a jail cell. If this model was used in 2021, it would still provide us with the same results, but our definition of what a cell is has changed several times over the years, and this prediction would no longer be useful to us.

Identifying model degradation

1) **Checking Model Predictions:** A common and simple method of identifying data degradation is by simply monitoring predictions and checking their accuracy. Having a pre-defined criterion of what will constitute as a correct model and using it to grade new model predictions will help track the degradation of data.
2) **Probability Estimations:** Probability estimations are used to measure the accuracy of a data model. Tracking the accuracy of a model and how it degrades over time is important. Long term trends can be identified using this method

3) **Checking Inputs for Data Drift:** This method is the most effective for data degradation. Reviewing data types, rages, and descriptive statistics will allow administrators to notice any changes in data definitions and importantly, they can tell if a model is outdated or not. Also defining a metrics with thresholds to track the difference between the accuracy of the training data and the data that is currently being used will lead to a strong indicator of model drift and degradation.
4) **Checking Concept Drifts:** Data drift distribution analysis can be used to analyze the drift between the current model and the training model. This method allows us to update model attributes that can be affected by concept drifts. Attributes such as marital status, where its values can change from “single” and “marries” can now include widows and those who are divorced. (Laurent, 2020)

Plan Maintenance

Maintaining a data model can be a tedious task, but on that is necessary.

- Usually, administrators will maintain both the logical and physical data model.
- Data model versions will be used to update the data model.
- Each version will compare the current version with previous versions to check for the two types of data drifts

Other maintenance activities include:

1. Creating subject areas and adding the relevant entities to those subject areas
2. Aligning subject areas
3. Validating versions
4. Testing new the new model. (learndatamodeling.com, 2015)

Conclusion

For this project we basically had to take the data of an organization, analyze it, clean it and build a model that would be able to give them valuable information about employees who earn more then R50k. We first cleaned the data, then we prepared it so it could ready for us to build a model, then we build a model, evaluated our model and finally deployed.

The following are our data mining result:

- Employees with a salary type of average did not earn more then R50k.
- There is a close correlation between the number of years an employee has worked and them earning over R50k.
- There is a close correlation between the department an employee works and whether they earn above R50k or not.

References

Laurent, B. (2020, July 20). *Model Monitoring Best Practices*. Retrieved May 13, 2021, from dominodatalab.com: https://www.dominodatalab.com/blog/model-monitoring-best-practices-maintaining-data-science-at-scale/

learndatamodeling.com. (2015, June 16). *Steps to create data models*. Retrieved May 13, 2021, from Learndatamodeling.com: https://learndatamodeling.com/blog/tag/maintain-data-model/#:~:text=Create%20subject%20areas%20and%20add,reports%20from%20the%20data%20model.