Team 9: Analytical Report on Joined Assignment

Abdallah Bazzan, Carolin Kressel, Cindy Moreno,

Janit Bilve, Tsholofelo Matabane, Vijay Yerramsetti

Hult International Business School

Data Analytics and Python, by Prof. Luis Escamilla and Prof. Chase Kusterer

MBAN – 2021/22

Word Count: 777

Executive Summary

We were able to develop a Naïve Bayes model that makes predictions about a customer's annual income given a set of characteristics from the U.S. census with a recall metric of 80%. Our model performed well with the test set since our metrics increased by up to two percent. Our journey as a team from python to the final report is overall a success, even though it has been a challenge for all of us.

Model Analysis

Upon training our model, we were able to predict whether a person's income is above or below 50k with a recall metrics of 80%. When testing our randomly selected test data that consisted of 20% of the raw data, we were able to further increase our metrics by up to 2%.

However, we should not disregard the False Negative Rate as we can say that our model will not be able to correctly predict the annual income 19% of the time. The government should also be aware of the False Positive Rate of around 27% as this can lead to wrong expectations regarding the peoples spending budget.

We believe that our trained model would be similar to our current model when selecting a different randomized training set of data from the current U.S. census. However, the predictive variables might change over time which makes regular updates necessary to ensure continuous accuracy in the future.

Reflection on Naïve Bayes Classification Technique

The Naïve Bayes classification method can be used in businesses to find potential customers. For example, an excellent Bayes model could help Tesla notice potential customers who are willing to switch to a Tesla car according to their corresponding demographic characteristics. This can help Tesla locate its new retail stores, study the progress of existing stores and to find exact locations to build a new charging site. Another business case could be for a dietary and protein supplements company that wants to understand why customer renew their subscriptions or not. The company could incentivize a customer to stay through the shipment of free samples and vouchers for friends.

The main advantage of the Naïve Bayes technique is the chance to predict behaviors or characteristics with data that is often already available. As soon as the template has been created and understood it is relatively simple to adjust the data to update or come up with a new system for a completely different purpose.

When creating the trained model for the U.S. census we had an initial advantage, by starting with clean data from python. However, the path to the clean data was long as it was the first time for us to fully prepare data for further analysis. The creation of the bins and then the model was also difficult at times as the understanding for the Naïve Bays method only grew while working with the template.

However, the at first most difficult parts of learning of how to clean data in python and understanding the Naïve Bayer method ended up also being the greatest. Learning by doing is the best form of increasing knowledge and now looking back on the excel and python file we don't fully understand what was so difficult to begin with.

Team Performance

With this assignment every team member was able to improve their skills at their individual level. Technical skills that we learned were how to import and work on a DataFrame in Python as well as the understanding and application of the Naïve Bayes method. However, the learning level in these areas differed based on previous understandings which makes this the area of opportunity for the team to further improve. Soft skill learnings concerned the importance of initial equal understanding of the assignment and the realization that the further a course progresses the earlier the team needs to start communicating about preparation needs to be done before the teamwork can start.

Carolin took the initiative to start with the python script. Vijay reviewed the script and tested it with different datasets. Abdallah and Tsholofelo wrote the report. Janit worked on the Naïve Bayes excel and Cindy reviewed everything with a particular focus on spelling and grammar.

Even though this sounds like a very strict separation of tasks we ensured working together as much as possible and contributing with ideas, questions, and solutions. During the team reflection we realized that we could implement better strategies to increase our productivity such as making sure everyone knew the requirements of the assignment well before meeting, properly

managing our time such as starting earlier and taking into consideration other exams and deliverables. This will increase the team's productivity and result in a high performing team.

Appendix

Training DataSet:

Confusion Matrix				
		ACTUAL		
		>50K	<=50K	Totals
Predicted	>50K	3772	3889	7661
	<=50K	908	10755	11663
	Totals	4680	14644	19324

Recall	80.60%
Precision	49.24%
False Positive Rate	26.56%
False Negative Rate	19.40%
Specificity	92.21%
Accuracy	75.18%

Testing DataSet:

Confusion Matrix				
		ACTUAL		
		>50K	<=50K	Totals
PREDICTED	>50K	963	952	1915
	<=50K	207	2710	2917
	Totals	1170	3662	4832

Recall	82.31%
Precision	50.29%
False Positive Rate	26.00%
False Negative Rate	17.69%
Specificity	92.90%
Accuracy	76.01%