Mark M. Lowe
DATA 4950

CAPSTONE: Nurse Attrition Classification

For this project, I chose a nurse attrition dataset because healthcare is a passion of mine, and, coming out of the pandemic, most of America is in a nursing shortage. For a healthcare business, having a model that can accurately predict which nurses are likely to attrit could prove to be a very valuable tool. For this reason, I focused on building and analyzing the results of a relatively simple model (Logistic Regression) and a more complex model (XGBoost Classifier) to see if I could predict which nurses attritted based on historical data.

I first chose Logistic Regression because of the results of running my data through a python module known as 'LazyPredict'. This module takes in a dataset and quickly runs an ensemble of models over the data. Once all the models run, the module prints out an ordered list of which models have the best chance to yield accurate predictions. Based on the results with my nursing attrition data, logistic regression landed in the top 3. Also, logistic regression is a relatively simple model to explain to business owners. Logistic regression works by predicting the probability of a binary target variable (in this case 'Yes' or 'No' for nurse attrition) based on a number of dependent variables related to the target variable. The predicted probability is determined by taking the probability of success and dividing it by the probability of failure. Normally, if the resulting prediction is > 0.5, the model predicts 'YES', while the model predicts 'NO' when this threshold is not met. However, a data scientist can put their 'thumb on the scale' by changing the threshold based on the specific use case of the predictions. This 'explainability' is the main reason why I chose logistic regression. In machine learning / predictive modeling use cases, communication with business owners can be quite difficult due to the difficult mathematical concepts some of these techniques utilize. Therefore, logistic regression is used throughout the industry because of its ability to accurately predict complex problems while also allowing for parameter analysis / prediction results that can be understood by non data scientists (the category to which most business owners belong). The fact that this technique also placed in the top 3 of the 'LazyPredict' results was the cherry on top.

Next, I chose to use an XGBoost Classification model because XGBoost is a powerful model that is used all throughout the data science world, with some even calling it the industry standard for predicting complex problems. My thought was to compare this complex model that can be quite difficult to explain (as extreme gradient boosting of decision trees combines multiple higher level statistical concepts) to the simpler logistic regression model. It is a difficult task to succinctly describe how XGBoost models work, so, for simplicity's sake, XGBoost utilizes a gradient descent algorithm to determine the addition of generally weak decision trees to sum into collectively stronger decision tree. This type of classification model yields binary predictions, along the same lines as logistic regression, but the model is more computationally expensive than the latter and can be harder to explain how the predictions are generated to business owners. Therefore, if the XGBoost yields similar results to the logistic regression model, I would feel comfortable using the simpler logistic regression in production, allowing for easier model development, explanation, and scalability.

Mark M. Lowe
DATA 4950

For explaining these models to healthcare companies, I would focus on the probability aspect of classification, as most individuals have a basic understanding of odds. Also, I would showcase which dependent parameters were most important to influencing the model's predictions (with a special focus on any parameters that do not match up with industry bias). I would do my best to talk about the results of the model as a story. In this case, a nurse manager who is losing nurses while not being able to figure out which nurses to target extra efforts towards in the hopes of not having those targets attrit. I would bring in the nursing shortage as a main factor to why my model could prove valuable, because it is able to find patterns in the data that standard practices may miss. Also, I would be sure to talk about the metric I chose to maximize results for healthcare businesses. My optimized metric was minimizing false negatives (when the model predicts the nurse will NOT attrit, & the nurse DOES attrit). I chose this metric because, as a business owner, I would want to know which of my nurses is predicted to attrit so that I can spend extra effort (incentive programs, promotion opportunities, etc.) on those nurses if I deem them valuable enough to keep. However, if the model misses these nurses entirely because it predicted 'will not attrit', a business owner does not even have a chance to narrow on these nurses to try to keep them. Nevertheless, this metric still needs to be balanced against overall accuracy. For instance, I could build a model that always predicts a nurse will attrit, thereby ensuring there are no 'false negatives'. However, these predictions will be basically worthless for the business as the model would give no insight into who to specifically target. Therefore, a happy medium needs to be struck between accuracy and false positives. This way, the business owners can trust the model to be correct a large majority of the time (>85%-90% accuracy) while also optimizing for false negatives. The tradeoff, in this case, would be 'false positive' (where the model predicts a nurse 'WILL attrit' but the nurse does not) predictions increasing. However, I don't necessarily think this is a bad thing. Some of these 'false positive' nurses may be right at the cusp of attritting, and, because the model predicted they 'WILL attrit', the business may spend some extra effort trying to keep them. This extra effort could lead to improved morale and longevity with the company, resulting in the model helping the business even in cases where its predictions were incorrect.

Overall, I would choose to use the logistic regression model to solve this classification problem. It's performance matched and, in some runs, even exceeded XGBoost's, and this model is easier to explain, develop, and scale to meet business needs. However, I would not say that my model is complete. My next steps would involve the business' subject matter experts on nurse attrition to ensure that my model's results match up with their domain knowledge. Also, I would be sure to bring in the business owners to create the optimal 'false negative / accuracy / false positive' relationship based on the business' resources and immediate needs. Lastly, I would search for more data while also looking to feature engineer new parameters that subject matter experts recommend for improving model performance. This model was trained on a relatively small amount of data (<2000 nurses) so bringing in more data would help myself and the business feel better about capturing more patterns into nurse attrition that the original dataset could have missed. In a perfect world, building in a timeseries aspect to the model so the business could know whether a nurse will attrit in 1 month, 3 months, or 6 months could also help the business better prioritize their resources to retain nursing talent.
Thanks for a great semester!!