Capstone Project

Executive summary

Predicting employee turnover in Archipelago International

Domonkos Surman

ACADEMIC YEAR | 2018-19
CAMPUS | Budapest
COURSE | MSc in Business Analytics
SUPERVISOR | Prof. Gabor Bekes
SUBMISSION DATE | 18 June 2019

Quest Hotel

favehotel
by aston

The Alana
HOTEL & CONVENTION CENTER - YOGYAKARTA
BY ASTON

Royal Kamuela

ASTON

HOTEL
NEO

Quest Vibe
★★
BY ASTON

NH
NOMAD HOSTELS

Kamuela
by Aston

Royal Alana

Alana

HARPER BY ASTON

Acknowledgements

External

I would like to thank Andras Kelemen and Archipelago’s Internal Team for all the assistance and support in integrating me to the company and providing the necessary resources for my Capstone Project.

Internal

I would like to thank my supervisor, Prof. Gabor Bekes for supporting me throughout all stages of my Capstone Project and for providing useful tips and best practices.

Executive Summary

About the project
As the final stage of my master's studies on the Business Analytics track at CEU, I did a research on employee turnover at Archipelago International. This company is a major player in the hotel management industry in South-East Asia.
The aim of the project was to prove that I am capable of doing an end-to-end data science project, using theoretical knowledge and technical skills learnt during the program as well as creating valuable insights together with long-term strategic suggestions.

Archipelago International is Indonesia's largest operator of hotels with 8 core brands with over 18,500 rooms and apartments and with over 130 thousand employees. With over 136 locations across Indonesia, the Philippines and Malaysia the company is trying to use the data that is gather during their operations and trying to implement it into their decision-making process. The data of current and past employees made it possible to execute an end to end data science project to predict employee turnover with the aim of predicting the expected length of employment of new hires.

What is employee turnover and why is it important?
When employees leave a company and someone has to take their place, that's called turnover. A certain amount of turnover/churn/attrition (all synonyms) is unavoidable, but too many employees leaving in any given period of time can give a hard time to the company. Turnover is expensive too, with total costs as high as one annual salary to locate and hire a replacement. For example, the costs of replacing a waiter or room maid is different than finding the next CTO.

Total costs include

- Cost of hiring
- Cost of onboarding and training
- Cost of learning and development
- Cost of time with unfilled role

Using the CRISP-DM methodology as my guide for the project the first step was to understand the business objectives which are

- to predict if an employee is at risk or leaving and
- to predict if a new employee is about to leave in a given timeframe.

An accurate prediction model can flag employees at risk of leaving resulting in management actions to keep them which at the end of the day can push down costs and increase productivity which can lead to higher margins. And profit of course, the ultimate goal for all business-oriented organization.

The data came from an AWS Athena connection from various departments which data I had to check and filter before I could merge them into one data frame that I used afterwards. After the initial cleaning the basic descriptive about the data showed the following

- 53.4% of the people who worked for Archipelago are currently active
- 20 variables that had no missing values

- Average employment duration – 27/34.5/18.6 months (all-time/currently active/terminated)
- Average age 30.5 / 29.6 years (currently active/terminated)
- The hotel sizes range from 1 to 348 rooms

To meet the business objective, I could use one already existing variable (*isterminated*) which had 2 values: 0 if the employee is still active in the company and 1 if the employee was already terminated.
To predict if a new employee is going to be terminated in a given timeframe new variables were needed which could be calculated from the previously mentioned *isterminated* variable and the length of experience variable (isterm_6/9/12). It was also important to drop all variables that has almost 0 variation in them together with the variables that correlated with each other. After all the cleaning and preparation, I had 20 variables and almost 30 thousand observation.

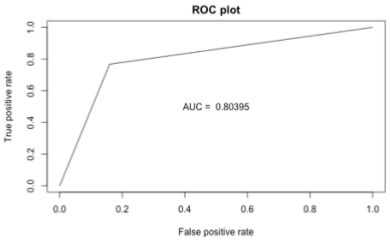
As all the target variables are binary, I chose to use different classification models: Logistic Regression (Logit), Random Forest and XGBoost.

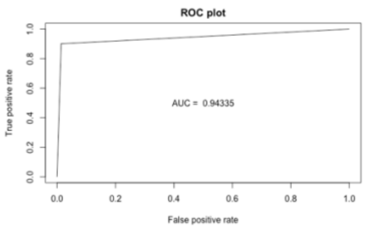
Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.
Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.
XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.

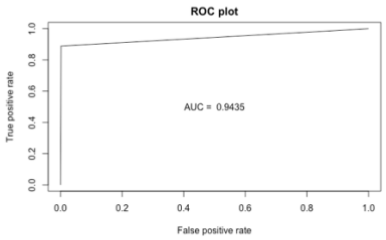
To measure the model performances, I calculated confusion matrixes and ROC/AUC values.

A confusion matrix is a table that is usually used to describe and measure the performance of a classification model (or “classifier”) on a dataset where the true values are known. In columns we have actual Positives and Negatives whereas in rows we have predicted positives and negatives.
True Positives and True Negatives indicate the cases when the prediction of the model is correct. False positives and false negatives are the opposite and considered as wrong predictions by the model.[1]
The following measures can be calculated from the confusion matrix:
1. Accuracy: (TruePoz + TrueNeg) / (TruePoz + FN + FalsePoz + TrueNeg)
2. Sensitivity - (TruePoz) / (TruePoz + FalseNeg)
3. Specificity - (TrueNeg)/ (FalsePoz + TrueNeg)

The ROC is a graph to show the connection/trade-off between sensitivity and specificity. AUC is defined as the Area Under the ROC Curve. Higher AUC value means better performance of the model
The Logit model performed the least good AUC = 0.804, while the RF / XGB models performed over 90% - AUC = 0.943 / 0.944





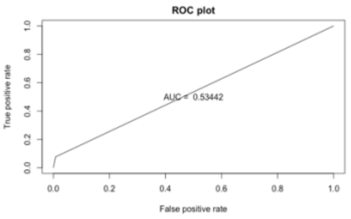


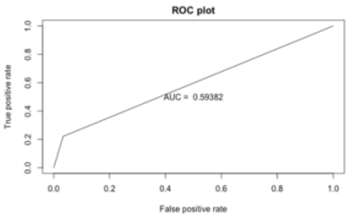
Looking at the confusion matrices it is visible that the in all three models False Negative predictions outweigh False Positive ones which short term can be beneficial for the employees who are aimed by the management's employee retention actions but as they were not the one who intended to leave long term this is not sustainable.

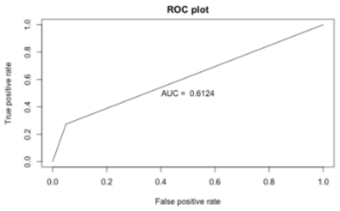
Next, I modelled if I can predict if a new employee is likely to stay for 6 – 9 – 12 months. I used the best performing XGBoost model for the analysis for 2 reasons: (1) XGB performed best previously (2) it was the fastest model to run.

Unfortunately predicting involving time didn't turn out to be too effective. The AUC values came out as the following:

- 6 months + terminated – AUC = 0.534 – no better than a coin-toss
- 9 months + terminated – AUC = 0.594 – slightly better chances than a coin-toss
- 12 months + terminated – AUC = 0.612 – best among the 3 models but too weak for further usage







Conclusion

Based on the currently available data we can predict with high accuracy which employee is likely to leave but we have problem predicting the timeframe.

Looking at the data source and types we can see that the data lack ordinal variables like salary information, job satisfaction on a scale, information on distance from home, performance index, number of holidays, years with current management, time since promotion etc.

Introducing new metrics to monitor performance throughout the company would help to finetune the already well performing model and hopefully would help with predicting when new hires are likely to leave.

In our fast-paced world employees become a major asset in every company's life and making extra efforts to keep them is worth pondering.

As a next step I would advise to evaluate the possibility of introducing new metrics and adjusting the previously discussed models to it.

Also, an inhouse evaluation on employee value would be needed to see if retaining employees really has extra value or in the hotel industry in Indonesia it is different from the rest of the world.