FINAL PROJECT ANALYSIS

Team 4: Prudhvi Raju Addada, Allison C Dibble, Sai Koushik Kollepalli, Srinivasa Satwik Mamidi Setti

DSCI 5240.009 Data Mining & Machine Learning for Business

November 13, 2023

Contribution Summary

Allison: Created summaries, formatted data descriptions & analyzed models to justify conclusions.

Sai: Analyzed models, formulated examples to provide business implication results.

Satwik: Developed models and analyzed results for findings.

Prudhvi: Developed models and analyzed results for findings.

Executive Summary

Attrition, also known as employee turnover, is an important metric for a company to keep at a minimum due to cost implications, loss of experienced employees, poor employee investment and customer relations.

The cost of attrition is considerable, encompassing the recruitment, onboarding, and training of new staff. The dataset was comprehensive and complete, with no missing values.

The models utilized were Logistic Regression, Decision Trees, and Neural Networks, each contributing to understanding factors affecting attrition. Logistic regression was used to determine the relationship between independent variables and the probability of the outcome of attrition. Decision trees helped figure out which employee has the highest probability to leave. Neural networks confirmed accuracy of the model using the confusion matrix.

Key recommendations inferred from the models are to reduce high attrition, include positive work-life balance, workload management and proper resource allocation.

Project Motivation/Background

- The cost implications of attrition are significant because it requires abundant resources to recruit, onboard and train new employees.
- Companies invest in the training and development of their employees.
- The loss of experienced employees can result in decreased productivity, potentially higher error and decrease in institutional knowledge.
- In businesses, customers may become frustrated if they repeatedly deal with new representatives, impacting the quality of service and the customer experience.
- The goal of this project is to work with a data set to figure out solutions to minimize attrition which can be beneficial to a company.

Data Description

The data set is about attrition of a company. This data set made by IBM has data of 1470 employees.

From Stat Explorer, assuming the acceptable values for skewness are –2 to +2, so none of the variables cross this acceptable region and do not have to be transformed.

From the summary statistics we can say that there are no missing values as the column for missing values says 0.

From the bar graph we can say that there are 237 employees who have left the company and 1233 employees who are still working in the company. Employees left the company is 1 and employees staying is 0.

Data Preparation Activities

- The data that we chose to work with was clean, so no additional work was done.
- Attrition is the employee status categorized into a binary target variable.
- No transformation required from the skewness.
- Put in a data partition node to split the data into training, validating and testing; 70% of data for training, 15% for validation to fine tune the model and the remaining 15% is used to test the model.

Models Used - Logistic Regression

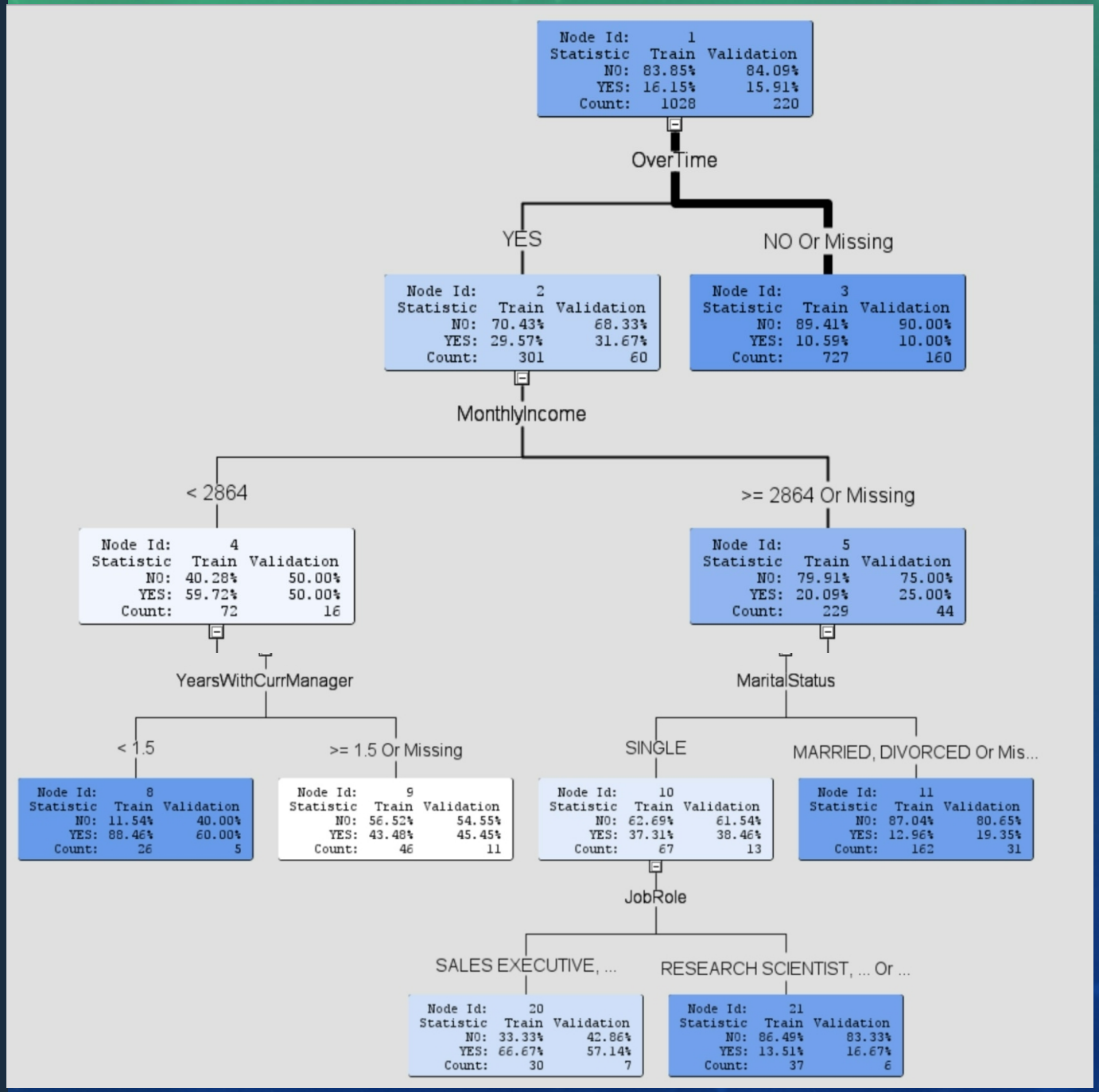
- Variable with highest odds is BusinessTravel.
- Variable with lowest odds is OverTime.
- Log(p/1-p) = e^5.2682+0.0465*1 = 1.048
- Increase in DistanceFromHome, odds of attrition increases.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	5.2682	0.8439	38.98	<.0001		194.070
BusinessTravel	Non-Travel	1	-1.0902	0.3041	12.85	0.0003		0.336
BusinessTravel	Travel_Frequently	1	1.1098	0.2011	30.46	<.0001		3.034
DistanceFromHome		1	0.0465	0.0119	15.19	<.0001	0.2092	1.048
EnvironmentSatisfaction		1	-0.3361	0.0928	13.11	0.0003	-0.2009	0.715
JobInvolvement		1	-0.7046	0.1367	26.55	<.0001	-0.2826	0.494
JobSatisfaction		1	-0.3586	0.0912	15.45	<.0001	-0.2198	0.699
MaritalStatus	Divorced	1	-0.6404	0.1936	10.94	0.0009		0.527
MaritalStatus	Married	1	-0.0987	0.1458	0.46	0.4984		0.906
NumCompaniesWorked		1	0.1349	0.0415	10.57	0.0012	0.1898	1.144
OverTime	No	1	-0.7885	0.1043	57.14	<.0001		0.455
TrainingTimesLastYear		1	-0.2218	0.0847	6.85	0.0089	-0.1568	0.801
VAR1		1	-0.0695	0.0133	27.14	<.0001	-0.3556	0.933
WorkLifeBalance		1	-0.3033	0.1376	4.86	0.0275	-0.1183	0.738
YearsInCurrentRole		1	-0.1907	0.0403	22.43	<.0001	-0.3764	0.826
YearsSinceLastPromotion		1	0.1484	0.0393	14.22	0.0002	0.2615	1.160

Models Used – Decision Trees

- The side with most observations is where the lines are the darkest.
- The weight of the line is heavier towards the ‘No Node’ to the right which means the number of observations is higher.
- The darker the node is the purer it is.
- The first split is overtime work.



Models Used – Neural Network

- Accuracy of the validation data is around 87.72%.
- Results found by adding the true positive and true negative and dividing it by the entire observations.
- Misclassification is 12.27%.
- Out of 100, this neural network predicts that an employee's attrition is correct 88 times.

Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=' '

False Negative	True Negative	False Positive	True Positive
73	836	26	93

Data Role=VALIDATE Target=Attrition Target Label=' '

False Negative	True Negative	False Positive	True Positive
20	178	7	15

Findings

- From logistic regression, employees who travel more are most likely to leave the company.
- Overtime has the lowest estimate, which leads to more retention if there is less overtime.
- Business travel has the greatest impact on attrition.
- From decision tree, SalesExecutive with monthly income less than $2,864 and YearsWithCurrManager greater than 1.5 show highest probability for leaving the company.
- Accuracy of neural networks, predicts attrition 88% of the time.

Managerial / Business Implications

- Setting realistic goals that are achievable by employees, must be implemented on the sales executives first because their group has the highest rate of attrition.
- From logistic regression, the company must reduce business travel by embracing more virtual meetings.
- Overtime is a huge reason for attrition and can be reduced by going through time management training and reviewing overtime policies.
- Promoting within an organization is key to retaining employees for a longer period. It reduces 2 of the highest attrition variables, YearsSinceLastPromotion and NumberOfCompaniesWorkedIn.
- Employees working with the same manager for more than 1.5 years have a high probability of leaving so shifting employees and managers might be good change that reduces this probability.
- Improving work-life balance and a culture of taking breaks can be implemented for a more pleasurable working experience.

Conclusions

- The model is good for the accuracy being 88%. Despite the high accuracy, there is still room for improvement in reducing this misclassification rate of 12%.
- The models give us information on what variables to focus on to improve attrition.
- Focusing on promotions within the company is a great practice to reduce retention.
- Overtime, YearsSinceLastPromotion, BusinessTravel, EnvironmentalSatisfaction are the highest variables for attrition.
- Employees with the highest attrition are Sales Executives and employees working with the same manager for more than 1.5 years.

References

- Shahid, A., Jain, R., Saud, S., & Ramirez, J. (n.d.). IBM HR Analytics Employee Attrition & Performance. https://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html
- Subhash, P. (2017, March 31). *IBM HR Analytics Employee Attrition & Performance*. Kaggle. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset