The Great Retention

By: Shruti B, Porshea E & Mouhamadou D

Project Summary

We are a consulting firm that aims to help clients transform data into tools for business optimization.

Using the data provided by the HR Department of an MNC, we looked into analyzing different factors to predict and improve employee retention.

We set out to answer the following questions:

What factors are important to an employee?

Which factors can the company focus on to improve their employee retention?

Data Cleanup

Data Cleanup

- We used the MNC HR dataset from Kaggle.com
- We dropped the Employee ID columns and transformed the Satisfactory and evaluation columns to float
- We also encoded the department and salary columns.

Data Columns of Interest

- Satisfaction_level: Satisfaction level of employee in percentage. 100% or 1 is very satisfied. 0% or 0 is not satisfied.
- Last_evaluation: Time from last evaluation in years.
- Number_project: Number of projects an employee is working on.
- Average_montly_hours: Average hours worked by employees in the last 3 months.
- Time_spend_commuting: Time spent by my employee commuting to the office.
- Left: If the employee has left the company.

Data Models

Models

- Looked at several models:
  - Logistic Regression Model
  - Decision Tree Model
  - Gradient Boosted Tree Model
  - Random Forest Model
  - Gaussian Naive Bayes Model
  - Support Vector Machines
  - XGBOOST Model

Comparison of Models

Gaussian Naive Bayes

	precision	recall	f1-score	support
Stay	0.90	0.84	0.87	2857
Leave	0.57	0.71	0.63	893
accuracy			0.80	3750
macro avg	0.74	0.77	0.75	3750
weighted avg	0.82	0.80	0.81	3750

Logistic Regression

	precision	recall	f1-score	support
Stay	0.80	0.93	0.86	2857
Leave	0.53	0.24	0.33	893
accuracy			0.77	3750
macro avg	0.66	0.59	0.60	3750
weighted avg	0.73	0.77	0.73	3750

Support Vector Machine

	precision	recall	f1-score	support
Stay	0.97	0.98	0.98	2857
Leave	0.93	0.91	0.92	893
accuracy			0.96	3750
macro avg	0.95	0.95	0.95	3750
weighted avg	0.96	0.96	0.96	3750

Decision Tree

	precision	recall	f1-score	support
0	0.99	0.98	0.98	2864
1	0.94	0.96	0.95	886
accuracy			0.98	3750
macro avg	0.96	0.97	0.97	3750
weighted avg	0.98	0.98	0.98	3750

Gradient Boosted Tree

	precision	recall	f1-score	support
0	0.98	0.98	0.98	2864
1	0.94	0.93	0.94	886
accuracy			0.97	3750
macro avg	0.96	0.96	0.96	3750
weighted avg	0.97	0.97	0.97	3750

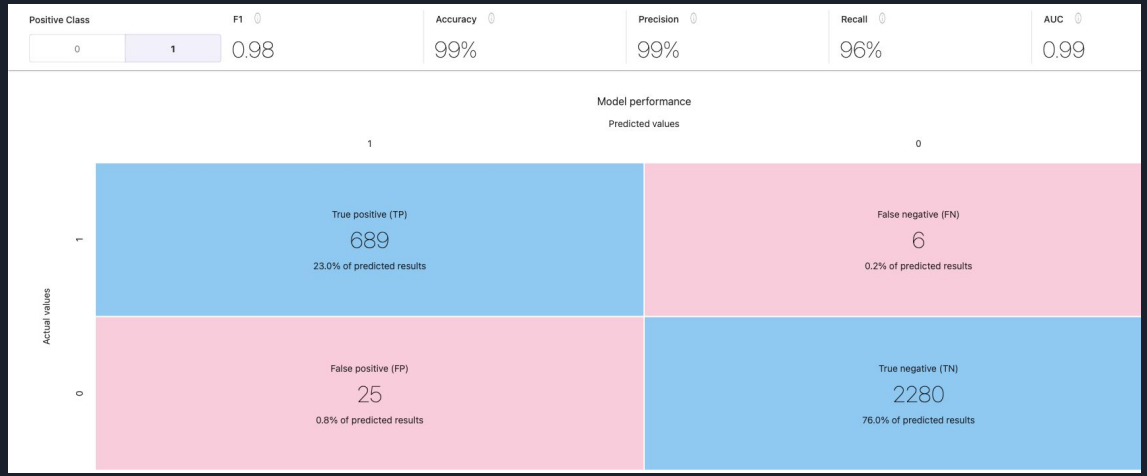
Random Forest

	precision	recall	f1-score	support
0	0.98	1.00	0.99	2864
1	0.98	0.95	0.96	886
accuracy			0.98	3750
macro avg	0.98	0.97	0.98	3750
weighted avg	0.98	0.98	0.98	3750

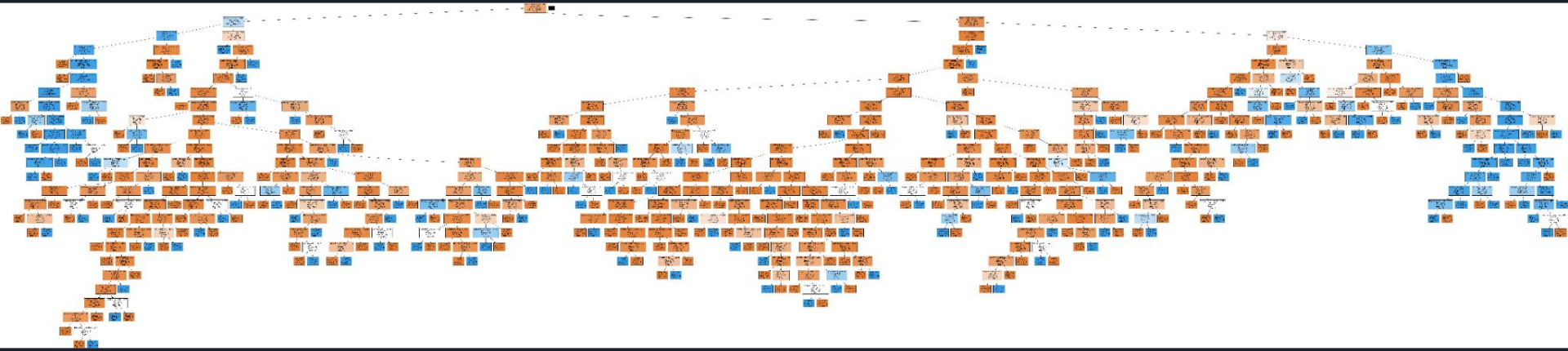
Comparison of Models Cont. - AWS XGBoost

Stay

Leave

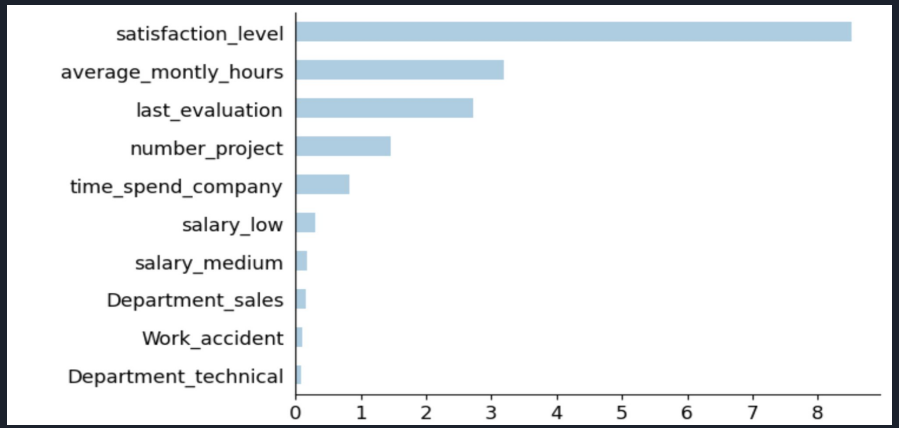


Decision Tree Model

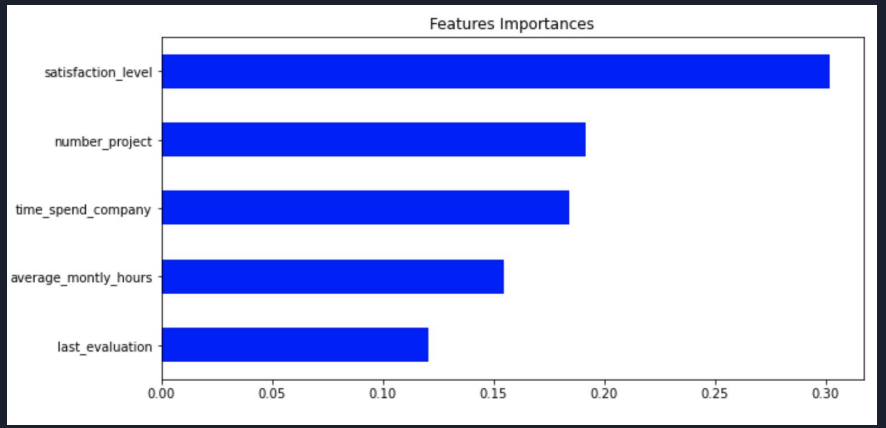


Best Model (Random Forest->Per our Coding XGBoost Model->Per AWS)

XGBoost Model-> Features-
99% Accuracy Score

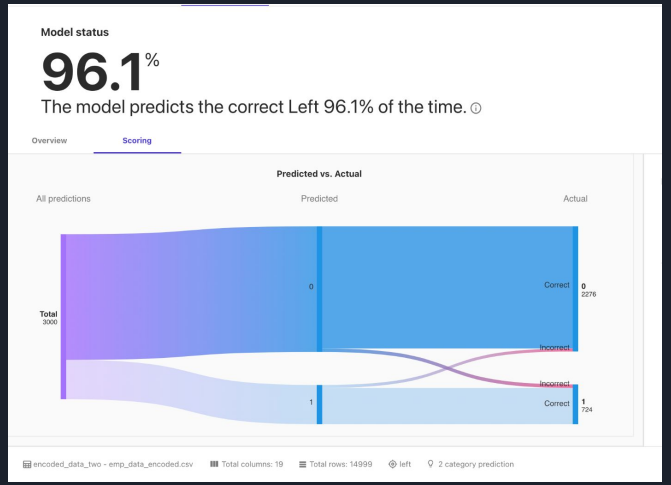


Random Forest Model-> Features-
98% Accuracy Score



What if we removed a couple of columns??

- Removed the Satisfaction Level and Last Evaluation Columns
  - How would this affect the models Accuracy score?



Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	2763	101
Actual 1	57	829

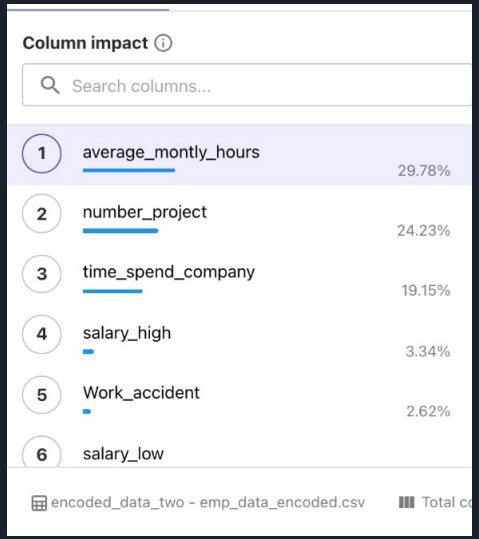
Accuracy Score : 0.9578666666666666

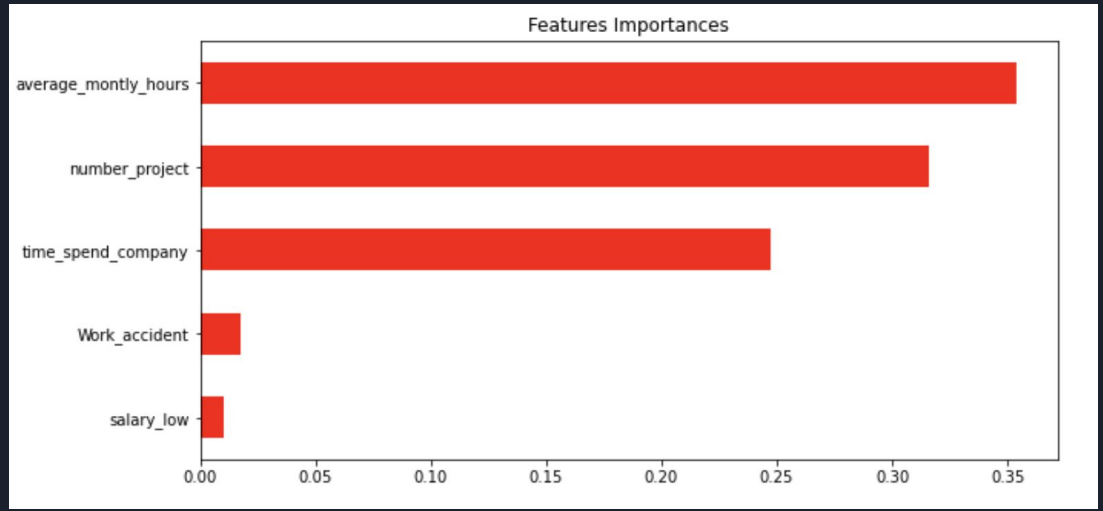
Classification Report

	precision	recall	f1-score	support
0	0.98	0.96	0.97	2864
1	0.89	0.94	0.91	886
accuracy			0.96	3750
macro avg	0.94	0.95	0.94	3750
weighted avg	0.96	0.96	0.96	3750

What if we removed a couple of columns??

- Removed the Satisfaction Level and Last Evaluation Columns
  - How would this affect the model importances Features?





Postmortem

- Complete shift in project early on
  - We started off wanting to analyze campaign finance and presidential elections, however there was insufficient data to create a model
- More time understanding how AWS Sagemaker works
  - Sagemaker can be very costly when running models
- Holiday in the middle of the project

- Next Steps
  - Test our model on similar datasets from other companies


Any
Questions