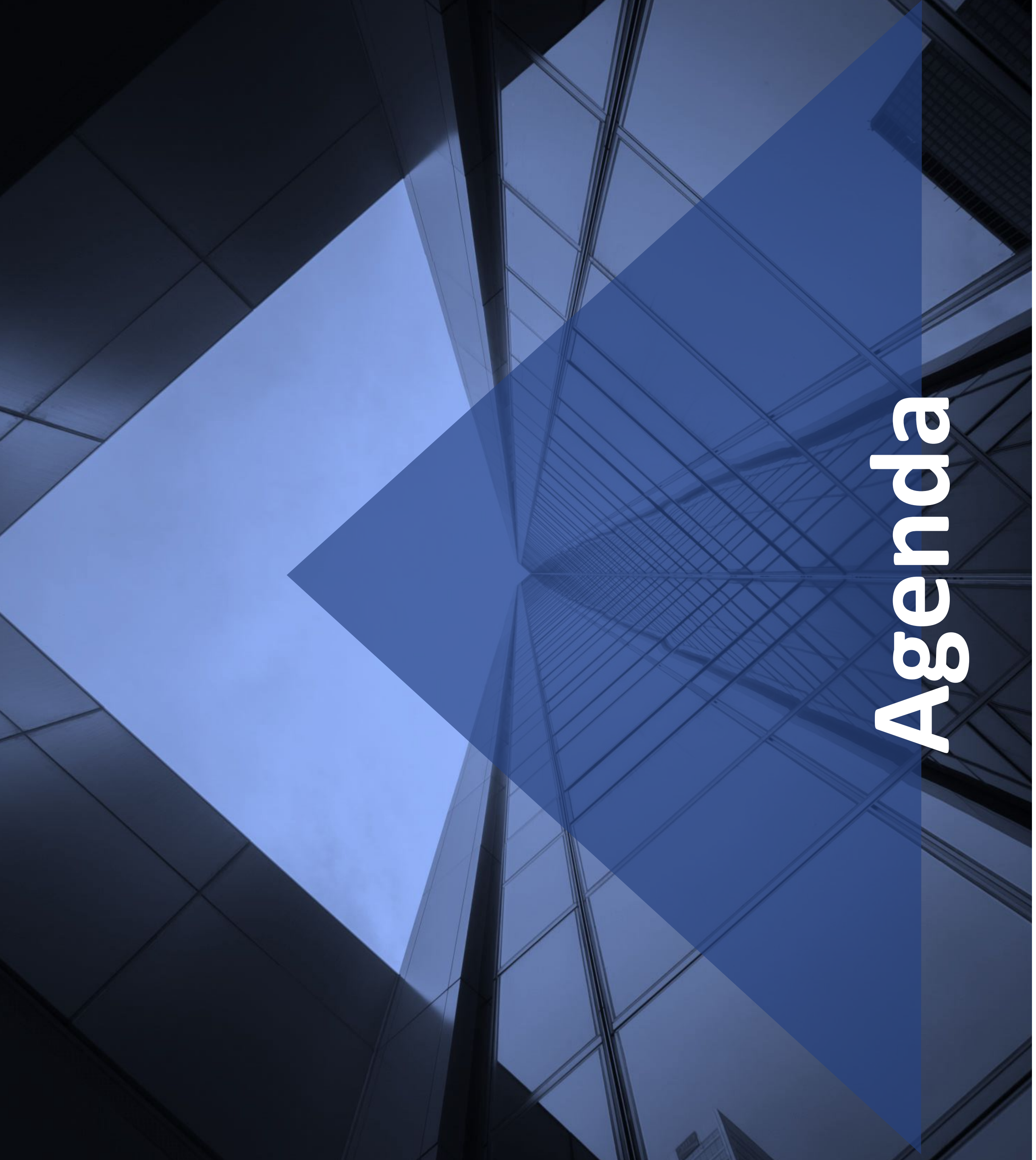
Classification Modeling to Predict and Reduce Employee Attrition

By Nicholas Bronson

MTS Shipping



Agenda

01 Problem & Key Questions

02 Methodology

03 Results

04 Conclusions & Recommendations

05 Next Steps

01
Problem & Key Questions

Problem:

Employee attrition has increased dramatically at MTS

Key Questions:

1) Who are the employees that are **most likely to resign?**
2) What factors have the **largest influence on employees departure**?
3) What might **appropriate and effective interventions** look like, and **what aspects might they target**?

02
Methodology

First Steps
- Problem scoping, defining criteria for success
- Data exploration
- Initial insights

Testing and Tuning
- Modeling & selection
- Model tuning and optimization

Results and Conclusions
- Conclusions and recommendations
- Outline for next steps

The Case for Classification & Key Metrics

*A classification model is well suited for MTS's problem as it allows us to **group employees** and **perform interventions on specific employees** who are at high risk of leaving.*

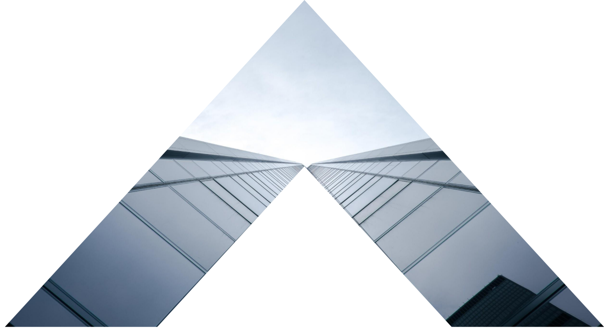
The Case for Classification & Key Metrics

*A classification model is well suited for MTS's problem as it allows us to **group employees** and **perform interventions on specific employees** who are at high risk of leaving.*

<u>Key Metrics</u>

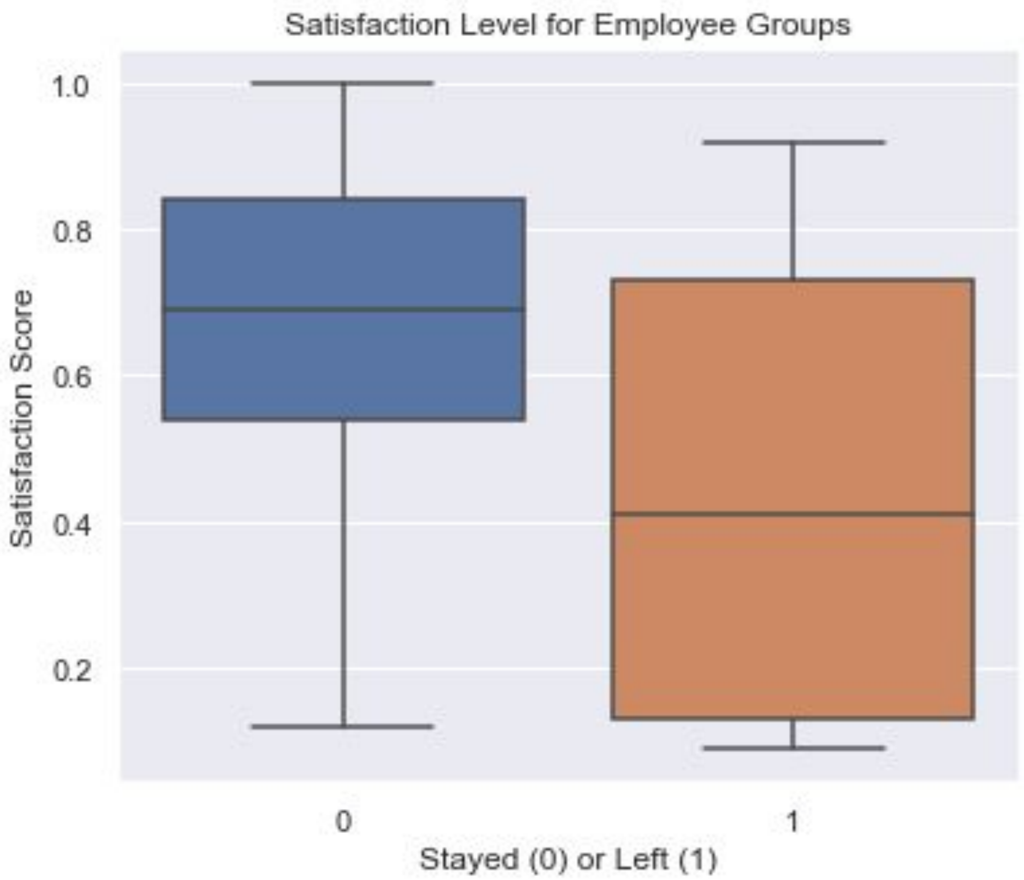
1. **Recall** - It is important to correctly identify employees that are likely to leave

1. **Accuracy** - This is secondary, however, the model should be able to segment as well as possible once recall is optimized



Initial Exploration

- Roughly **76%** of the employees in the dataset remain at MTS
- Satisfaction levels are higher on average for employees who stayed versus those who left
- Only **6.6%** of employees in the **high-salary** band have left the company
- There appears to be a relationship between **high working hours** and **low satisfaction scores**



03 Results

- Models tested included: K-NN, Logistic Regression, Decision Tree, Random Forest
- Feature engineering and optimization Included: creation of dummy variables, testing dummy variables vs. numeric substitution

Strong Performance Following Optimization

Random Forest:
- Recall: 97.3%
- Accuracy: 98.2%

Decision Tree
- Recall: 96.7%
- Accuracy: 97.9%

K-Nearest Neighbors (N=9)
- Recall: 92.7%
- Accuracy: 93.9%

Weaker Performance even after Optimization

Logistic Regression:
- Recall: 82.6%
- Accuracy: 71.8%

03 Results

- Models tested included: K-NN, Logistic Regression, Decision Tree, Random Forest
- Feature engineering and optimization Included: creation of dummy variables, testing dummy variables vs. numeric substitution

Strong Performance Following Optimization

Random Forest:
- Recall: 97.3%
- Accuracy: 98.2%

Decision Tree
- Recall: 96.7%
- Accuracy: 97.9%

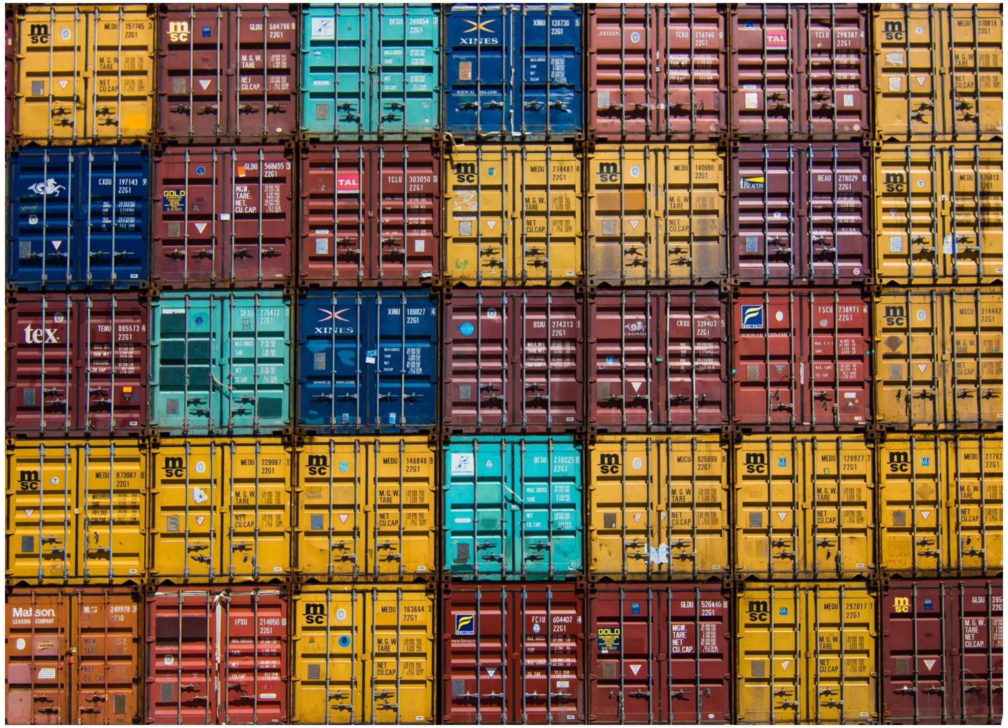
K-Nearest Neighbors (N=9)
- Recall: 92.7%
- Accuracy: 93.9%

Weaker Performance even after Optimization

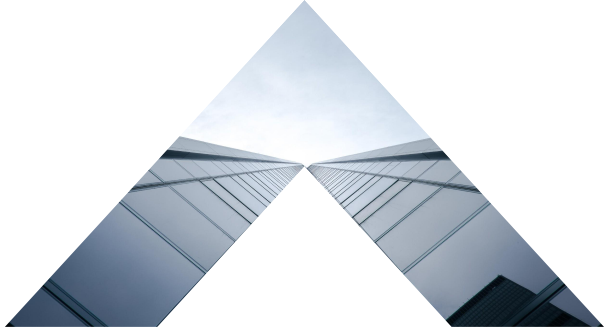
Logistic Regression:
- Recall: 82.6%
- Accuracy: 71.8%

Feature Importance

- A random forest regressor assisted in determining the following factors have the most influence on attrition:
  - Satisfaction level
  - Recency of last evaluation
  - Number of projects
  - Average monthly hours
  - Time spent with the company



04 Conclusions & Recommendations



Conclusions:

1. Our model can be used **now** and perhaps **again in several months** to classify employees as those likely to leave and those likely to stay

2. **Satisfaction, time spent with the company, and working hours** a month are features influencing employee retention

3. An intervention might target specific employees, those misclassified as employees who would leave, and those who are classified as so in the future

Recommendations:

1. Roll out an intervention **targeted at identified employees**, test intervention efficacy

2. Consider **adjusting policies around working hours, and incentives/recognition for longstanding employees** company-wide

3. Inquire about and address employee satisfaction levels, create a feedback loop through survey and response

05 Next Steps

1) Gather additional data, re-train models continuously
    a) More specific salary information, remote vs. in-person,
2) Additional ensembling and feature engineering
3) Explore Gradient Boosted Tree models
4) Measure efficacy of intervention methods, adjust as necessary

THANK YOU

MTS Shipping