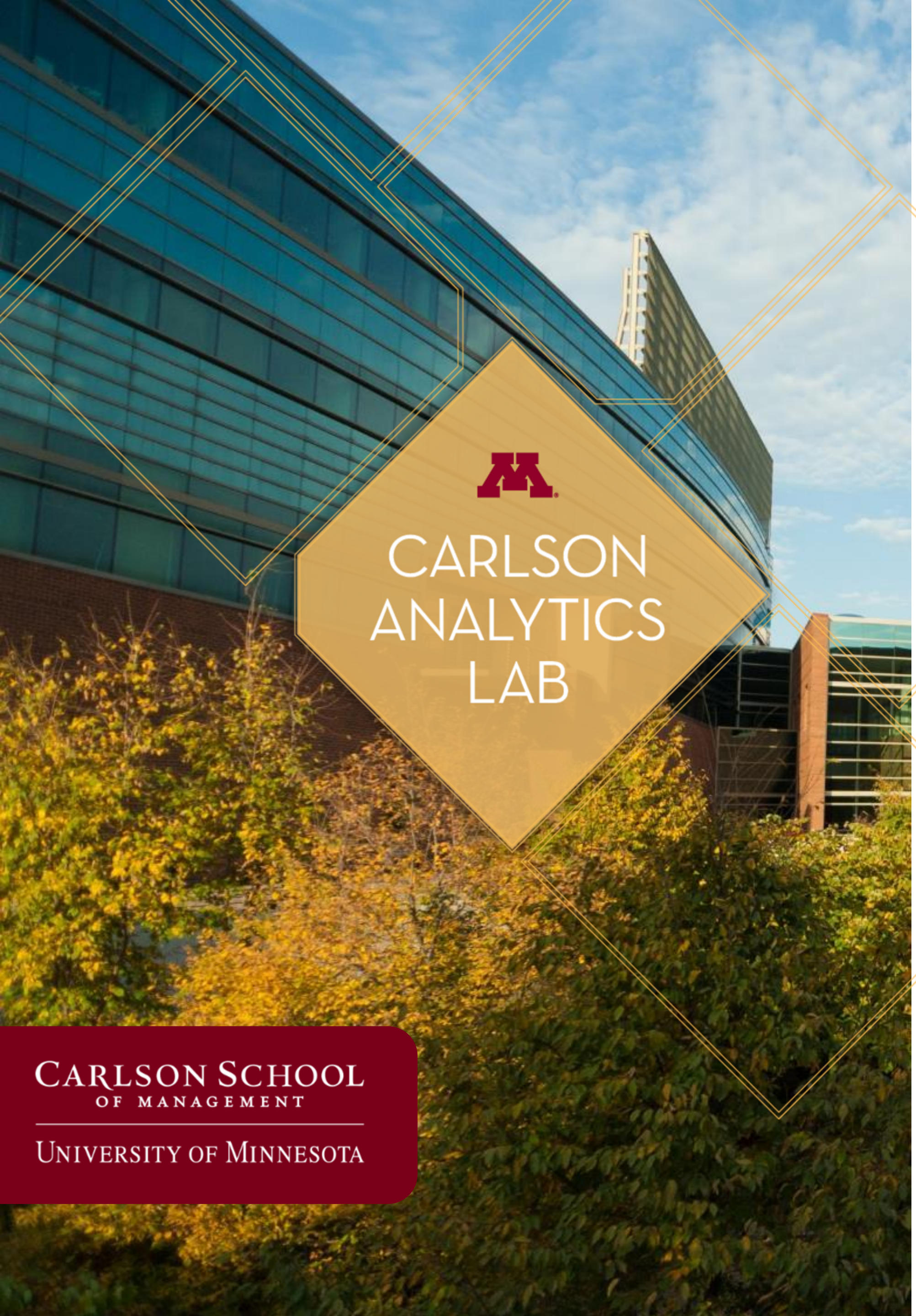


UMN OHR project

Predicting employee departures

Faculty/Staff Model Building & Final Presentation Prep

July 28, 2020

Kevin Grady, Wendy Lu, Anthony Meyers, Danny Moncada,
Claire Ryan & Jonathan Watkins

Today’s Agenda

This Week’s Progress

Next Steps

This Week's Progress

Table Generation & Feature Engineering

- Conversion of code to pre-packaged Python script - *in progress*

Model Building

- Faculty modeling - Classifier
    - Refined/tweaked parameters to determine best performing model with existing data and features
- Staff modeling - Classifier
    - Refined/tweaked parameters to determine best performing model with existing data and features
- Survival analysis
    - Continued development and testing of model outcomes

Setting up machine learning pipeline for Staff & Faculty Classification

A. Take output of “flattened” table generation and feature engineering pipeline (set of Python scripts) which is delivered as a “final” CSV file

B. Select a set of features to use for training the models (using our hypothesized feature set)

C. Transform the features into a format that is usable by the algorithms (see notes for details)

D. Split the data in training/testing sets (in our testing we used ⅔ for training, ⅓ for test)

*Normalization is the process of scaling individual data points to have a normal or standard distribution, and is critical because many machine learning estimators behave erratically if individual features do not look like normally distributed data.

** Many machine learning algorithms cannot operate on features that are “label” values vs. numeric values; a DecisionTree classifier **can** handle categorical labels directly but a Logistic Regression model requires numeric inputs and outputs.

Staff Classification & Modeling - Outcomes

- Machine learning pipeline created to classify professional staff employees using entire five year history
- Using data provided in VM, here was the best performing model using the LightGBM algorithm

- Top Five Feature Importances
  - Time at University of Minnesota
  - Pay
  - Supervisor Number of Reports
  - Time to Last Raise
  - Number of Raises

LightGBM – Model Evaluation

Confusion Matrix	
True Negatives	False Positives
False Negatives	True Positives

	Active Employees	Churned Employees
Active Employees	4115	102
Churned Employees	187	**783**

	Precision ***(TP / (TP + FP))***	**Recall** ***(TP / TP + FN)***	**F1** ***(2 * (Precision * Recall / Precision + Recall))***	**Support** **(how many testing examples were available)**
Active Employees (0)	0.953	0.983	0.968	4217
Churned Employees (1)	0.915	0.791	0.848	970

Faculty Classification & Modeling – Outcomes

- A similar pipeline was used to classify faculty
- As with Professional Staff, LightGBM was found to be the best performing predictive model using the entire history

- Top Five Feature Importances
  - Pay
  - Time at University of Minnesota
  - Supervisor Number of Reports
  - Department Size
  - Time to Last Raise

LightGBM – Model Evaluation

Confusion Matrix	
True Negatives	False Positives
False Negatives	True Positives

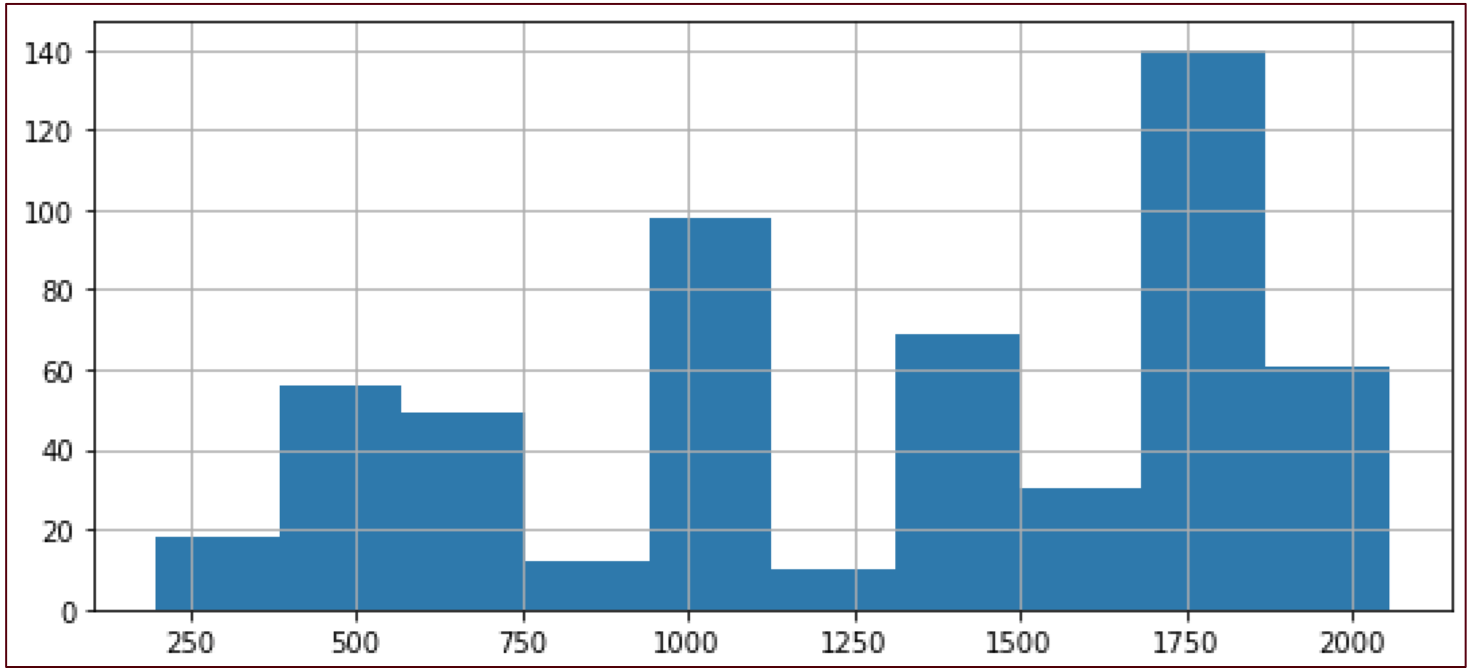
Faculty		
	Active Employees	Churned Employees
Active Employees	1334	25
Churned Employees	70	**164**

	Precision *(TP / (TP + FP))*	**Recall** *(TP / TP + FN)*	**F1** *(2 * (Precision * Recall / Precision + Recall))*	**Support** **(how many testing examples were available)**
Active Employees (0)	0.982	0.950	0.966	1359
Churned Employees (1)	0.868	0.701	0.775	234

Survival Analysis Modeling & Outcomes

Proportional Hazard (PH) Model Results

Histogram of Predicted Median Survival Times (in weeks)



Model Evaluation:
- Used holdout data to evaluate performance
- Achieved a concordance score of 0.85

Model Application:
- Predicted median survival times for non-resigned individuals

Model Results:

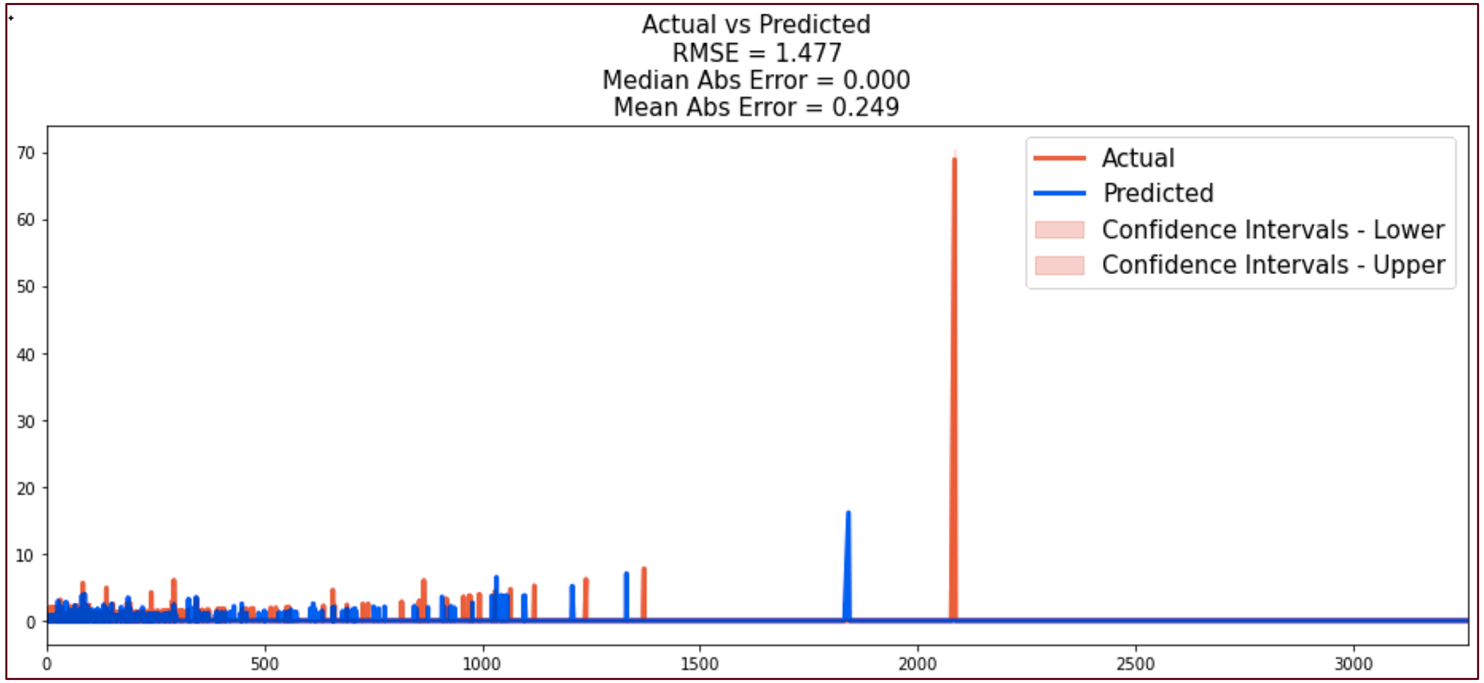
Mean:	1292	1st Quartile:	944
Min:	200	2nd Quartile:	1440
Max:	2055	3rd Quartile:	1773

Note: Durations given in weeks

Survival Analysis Modeling & Outcomes

Conditional Survival Forest (CSF) Model Results

Actual vs Predicted Survival Times



Note: Durations given in weeks

Model Features:

- Allows us to incorporate more data such as ZDEPTIDs and Locations
- Does not make certain assumptions

Model Evaluation:

- Used holdout data to evaluate performance
- Achieved a concordance score of 0.75

Model Application:

- Predicted survival times for non-resigned individuals

Challenges when dealing with running predictions with two+ month lead time

- Few instances of churn led to very poor precision scores
- Lacking key features to generate credible predictions
- Classifiers (on their own) cannot be relied upon for **credible** predictions with the limited data available

- Techniques used to overcome issue:
  - Oversampling/undersampling
  - Different prediction windows (3 months and 6 months)
  - Training / testing splits using multiple periods (TimeSeries splitting)

LightGBM – Model Evaluation on three month window

Professional Staff (3 month window)		
	Active Employees	Churned Employees
Active Employees	12263	212
Churned Employees	112	**3**

*Looking at the confusion matrix for model evaluation on a three month window, we see much less **support** for employee resignations; using a subset of data from December 2019 to March 2020, there were only 115 resignations, leading to a significant class imbalance.

	Precision ***(TP / (TP + FP))***	**Recall** ***(TP / TP + FN)***	**F1** ***(2 * (Precision * Recall / Precision + Recall))***	**Support** **(how many testing examples were available)**
Active Employees (0)	0.983	0.983	0.983	12475
Churned Employees (1)	0.014	0.026	0.018	115

Next steps

Next Steps

- Finalize machine learning pipeline and begin updating Knowledge Transfer document
- Provide OHR with lessons learned / next steps for improving model outcomes
- Prepare for final presentation & begin updating Knowledge Transfer