Project Mid-Semester Update

Identifying Key Variables to Explain Employee Performance

MGT 6203 | Fall 2024 | 11/2/2024

Ben Mascott (Analysis Co-lead), Erick Ordonez (Project Manager),
Jerry Fang (Data Visualization Lead), Isabella Reyna (Analysis Co-lead)

Georgia Tech

Recap of Project Objectives

Project Objectives

1. Determine **which factors**, such as years of experience, job satisfaction, and education level, **significantly impact employee performance** to help Human Resources (HR) tailor support and intervention efforts.
2. **Provide HR with a model to predict employee performance** so they can allocate their limited time and budget most effectively to support employees who are more likely to be low-performers.

Business Problem

Our team was hired by a firm whose HR department would like to improve employee performance, but has **limited resources** to do so. By identifying which variables best explain the variability in employee performance and developing a model to predict performance, we can provide HR with the tools they need **to save time and money** by **targeting resources** at variables that are associated with higher employee performance.

Data Collection Update

Data Source

The Employee Performance dataset (a pre-approved dataset) is the sole data source for this project.

Data Collection Status

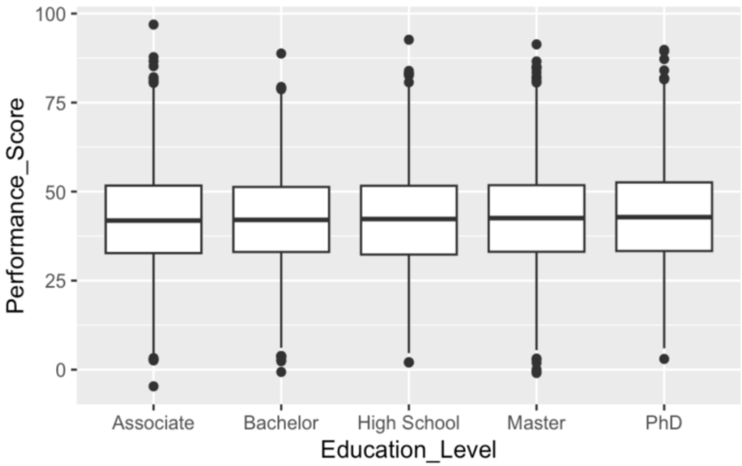
Data collection is complete as of early October 2024.

Data Challenges

We are using a pre-approved dataset and have not faced any challenges downloading the data and reading the data into R. The data does not have any missing values.

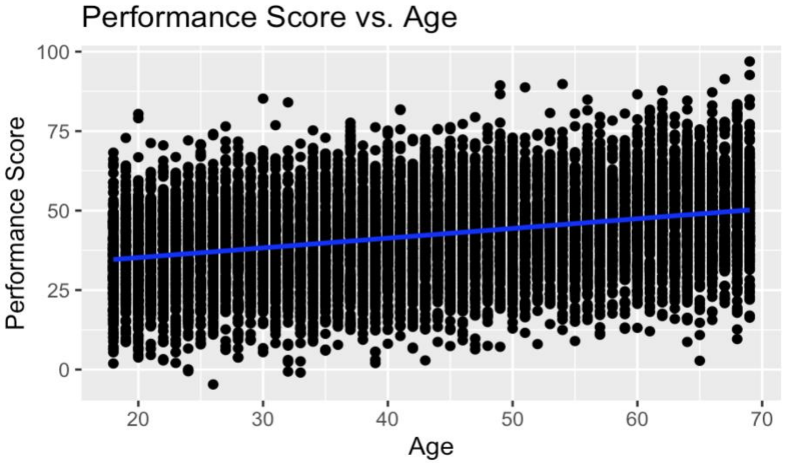
Exploratory Data Analysis (EDA) - Visualization example charts

Boxplots



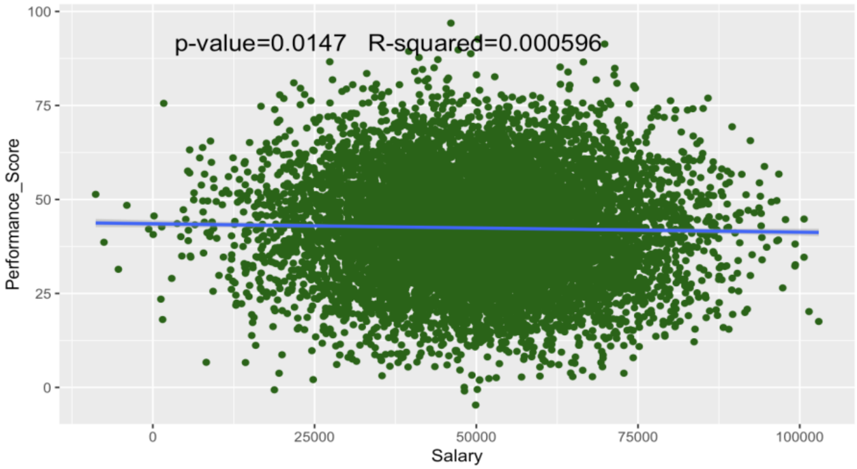
- The sample boxplot above displays a wider outlier range in Performance Score for Employees with an **Associate Degree**.
- The boxplots display slightly different interquartile ranges and medians of performance scores for each **Education Level**.

Scatterplots/Correlation Matrix



- The scatterplot above displays a strong positive linear correlation between **Performance Score** and **Age**
- The line of best fit (blue line) summarizes the linear relationship between **Age** and **Performance Score**
  - Its upward slope indicates a positive correlation, which means as Age increases, Performance Score also increases.

Improving R-Squared



- As represented in the chart above, the p-value=0.0147 and R-squared=0.000596 for the Salary predictor:
  - This indicates that **Salary** is statistically significant when alpha = 0.05. Note: This does NOT mean Salary will necessarily be significant in the full regression model with additional predictors present.

Plots were created for every predictor. No quantitative predictors will require a transformation. No missing data.

Analytical Approach and Methodology

Preliminary Results for Linear Regression Model with Stepwise Selection: Model assumptions hold for the model, and no transformations are needed for the response or predictors. Using a Cook's distance threshold of 1, no outliers were detected. Using a VIF threshold of 5, multicollinearity is not a concern.

Technique	Rationale
Data Split	We split the data into an 80/20% **train/test split** so we can assess model performance and compare different models. This helps ensure we are not providing the company with a model that overfits the training data.
Full Multiple Linear Regression Model	We fit a linear regression to the training set and verified the model assumptions using **residual plots** to assess goodness of fit. We also used **Cook's distance** to identify outliers and calculated variance inflation factors (**VIFs**) to identify multicollinearity.
Regression Model with Stepwise Selection (Forward, Backward, and Forward-Backward)	We used **variable selection techniques** to reduce the number of variables in the full regression model because along with predicting employee performance, the business problem also entails helping HR allocate limited resources more efficiently. By reducing the number of variables, HR can use the model to **allocate resources to key areas** and then when they want to collect data in the future to measure the results of their intervention, they don't need to waste time collecting data for every predictor. **Note:** Model assumptions, outliers, and multicollinearity were also assessed in this model. **Predictors selected with Stepwise techniques:** Experience, Work Hours Per Week, Age, Education Level, Annual Bonus, Job Satisfaction. **Note:** All three stepwise regression models selected the same variables.

Progress Towards Objectives

Milestones Achieved

✓ **Completed EDA**

- Box plots for each categorical predictor in the dataset; Scatter plots for each quantitative predictor.
- Correlation matrix for quantitative predictors.
- Linearity assumption assessed for quantitative predictors.

✓ **Completed Preliminary Modeling**

- Full multiple linear regression complete with diagnostic tests (test for assumptions, outliers, and multicollinearity)
- Stepwsie Regression models complete with diagnostic tests.

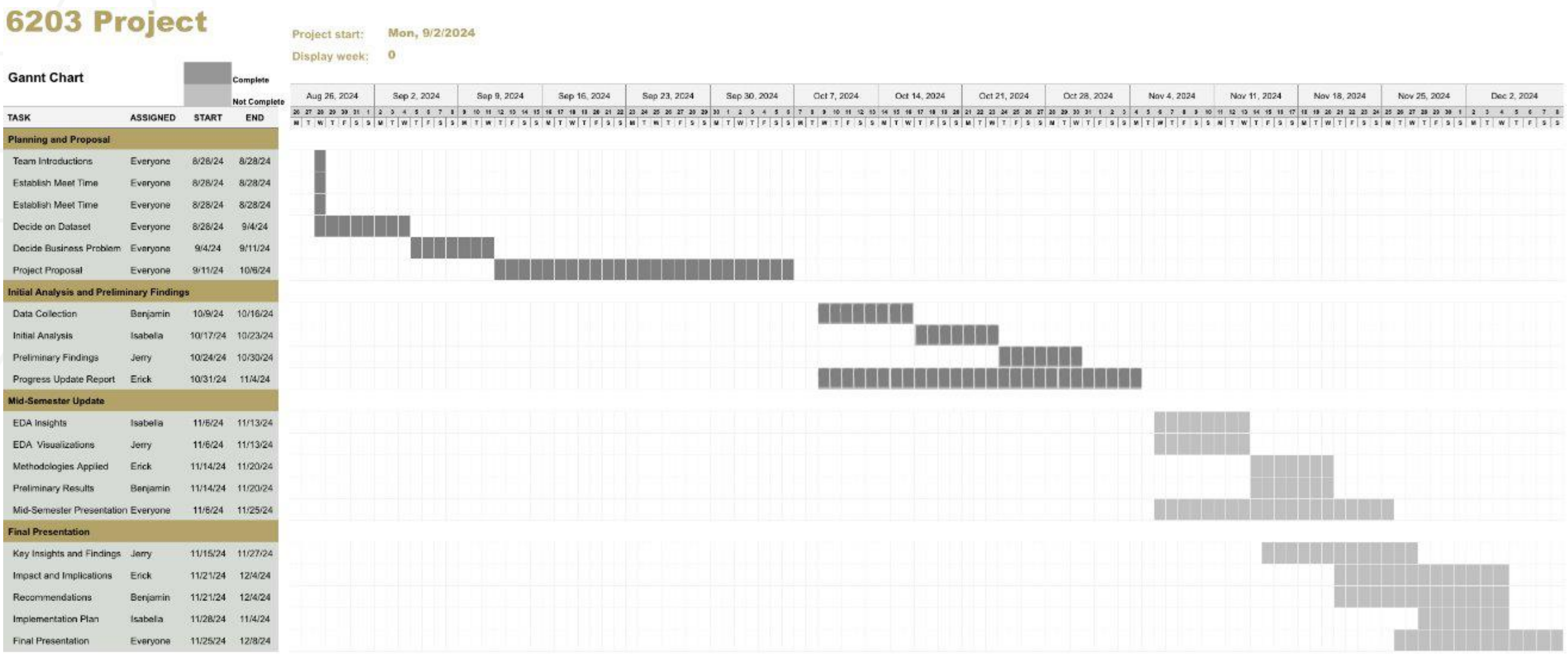
Alignment With Goals

Through exploratory data analysis, we gained a deeper understanding of our dataset and evaluated key linearity assumptions. Initial modeling has identified a subset of predictors that most effectively explain the variation in employee performance. This progress directly aligns with our goal of pinpointing the most impactful factors and developing a robust model to guide HR in supporting employee growth and productivity.

Adjustments Made

Based on the TA's feedback, we've refined our approach to focus on developing multiple models using various variable selection methods. Per the TA’s feedback, we’re now also working on fitting decision-tree and random forest models to the cleaned data. We'll compare the developed models by fitting them on a training set and evaluating their performance on a test set, using metrics such as R-squared and RMSE. We decided not to pursue separate models for different departments or education levels, as the data showed minimal variation between these groups, making this approach less effective.

Plan For Completion



6203 Project

Project start: Mon, 9/2/2024

Display week: 0

Gannt Chart

▨ Complete

□ Not Complete

TASK	ASSIGNED	START	END
Planning and Proposal			
Team Introductions	Everyone	8/28/24	8/28/24
Establish Meet Time	Everyone	8/28/24	8/28/24
Establish Meet Time	Everyone	8/28/24	8/28/24
Decide on Dataset	Everyone	8/28/24	9/4/24
Decide Business Problem	Everyone	9/4/24	9/11/24
Project Proposal	Everyone	9/11/24	10/6/24
Initial Analysis and Preliminary Findings			
Data Collection	Benjamin	10/9/24	10/16/24
Initial Analysis	Isabella	10/17/24	10/23/24
Preliminary Findings	Jerry	10/24/24	10/30/24
Progress Update Report	Erick	10/31/24	11/4/24
Mid-Semester Update			
EDA Insights	Isabella	11/6/24	11/13/24
EDA Visualizations	Jerry	11/6/24	11/13/24
Methodologies Applied	Erick	11/14/24	11/20/24
Preliminary Results	Benjamin	11/14/24	11/20/24
Mid-Semester Presentation	Everyone	11/6/24	11/25/24
Final Presentation			
Key Insights and Findings	Jerry	11/15/24	11/27/24
Impact and Implications	Erick	11/21/24	12/4/24
Recommendations	Benjamin	11/21/24	12/4/24
Implementation Plan	Isabella	11/28/24	11/4/24
Final Presentation	Everyone	11/25/24	12/8/24

Expected Challenges

Potential Obstacles	Mitigation Strategies
Model Assumptions. We needed to make sure that model assumptions were met for Multiple Linear Regression. (Linearity, Constant Variance, Normality, Independence).	We made sure no model assumptions were being violated by checking assumptions with residuals plots, QQ plots, and scatterplots. No assumptions were violated which means we are clear to use Multiple Linear Regression for our analysis.
Outliers. Outliers can be problematic specially if they are very influential. They make our fit biased and change the magnitude of our coefficients which can lead to incorrect conclusions.	To identify outliers, we used Cook's distance with a threshold of 1. No outliers were detected.
Multicollinearity. Correlation between independent variables could impact the statistical significance of some predictors.	To detect multicollinearity, we used the Variance Inflation Factor (VIF) technique with a threshold of 5, and found no instances of multicollinearity in models developed thus far.

Questions and Feedback

Area for Feedback #1: Recoding Categorical Predictors

- **Background:** The Regression Model developed using Backward Selection found some but not all the levels for the categorical predictor “Education Level” to be significant at the alpha=0.05 level.
- **Question:** Can we combine some of the levels for categorical predictors (such as “highschool” and “bachelors” since neither was significant) to decrease the number of predictors (specifically, dummy variables) in the model? If we combined predictors, would it be better to regroup the predictors using Group LASSO, or recode predictors directly within the training and test sets?

Area for Feedback #2: Comparing Models with Stepwise Regression vs Regularized Regression

- **Background**: We’ve performed an 80/20 train/test split of the data. We used the training set to fit the full regression model and plan to use the test set to evaluate the performance of different models.
- **Question:** Since we’ll have to scale the data before fitting LASSO or Elastic Net using the training set, do we also need scale the test set data or do we use the selected variables from LASSO and Elastic Net to refit new linear regression models on the unscaled and unseen test data in order to correctly compare model performance across different models?

Any additional feedback outside of the specific areas addressed above would be appreciated as well!

Summary and Thank You

Project Purpose

- Identify key variables associated with higher employee performance
- Develop a model to predict employee performance

Progress So Far

- Ensured assumptions for Multiple Linear Regression have been met
- Constructed a plan to assess the performance of our model (train various models on training set and evaluate with test set)
- Created visualizations (box plots, scatter plots, residual plots, correlation matrix) to better understand our features

Next Steps

- Apply non-linear models (Decision Tree and Random Forest)
- Develop additional linear regression models using more feature selection techniques (LASSO and Elastic Net).
- For each new model, evaluate performance on the test set and calculate R-squared and RMSE, then compare models and select the best one.