Predictive Revenue Using Machine Learning Methodology

Bryan Lee, Sanjana Shinde, Emily Weber

INTRODUCTION AND PROJECT BACKGROUND

Background
CLA is a professional services firm delivering integrated wealth advisory, outsourcing, audit, tax, and consulting services.

Project Goal
Help CLA improve its methodology for predicting recurring revenue streams for leadership and planning purposes

Importance
Predict revenue to understand how much reinvestment should be made and where in the company. In addition, our leaders will better be able to better track our progress.

RESEARCH METHODOLOGY

❖ Utilized RStudio to create train and test data to assess and describe CLA revenue streams.
❖ Used the CLA time and billing database along with dplyr/dbplyr to transform and extract data that could be used to assess relationship between time/wip and fees.
❖ Employed traditional forecasting (ARIMA, ets, etc.) time series models where applicable to forecast future (recurring) revenue
❖ Used other methods of forecasting such as random forests, boosting, and resampling based methods.

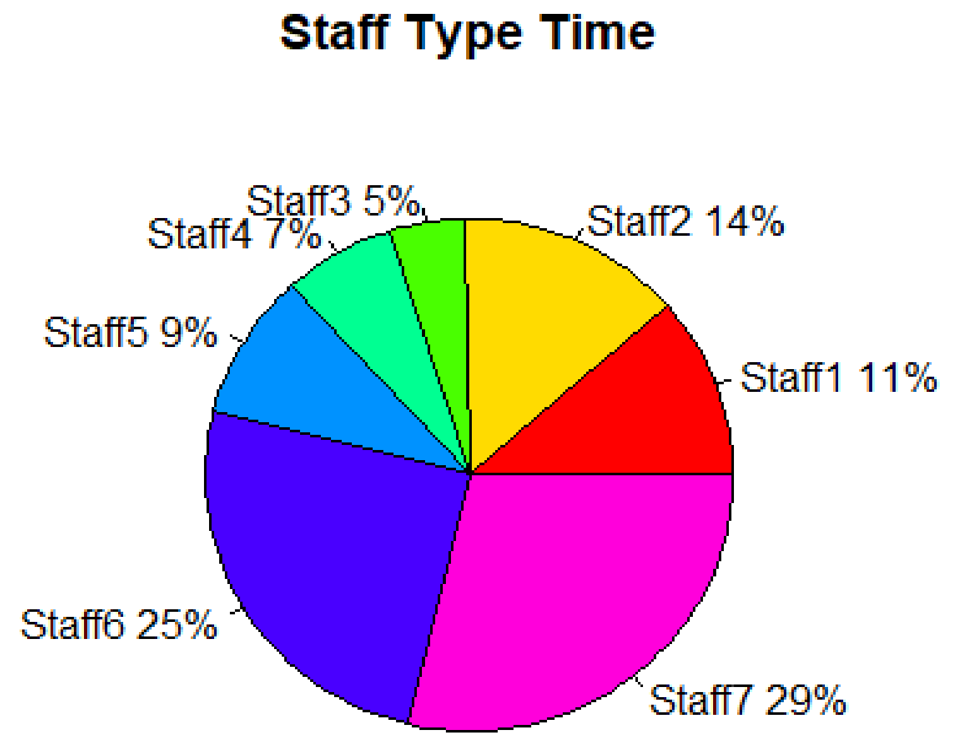

Figure 1:

Visual representation of the different staff types working on engagements for CLA.

CONCLUSION

❖ Using the CLA Time and Billing Database, along with dplyr/dbplyr, and tidyverse
❖ Quantify the accuracy of our results using RMSE
  ❖ The RMSE is valid because it tracks how "off" the points are by calculating the average square difference (squared residual) between the observed and expected; low RMSE means the prediction is accurate
❖ Time Series Model (Figure 2)
  ❖ Generally, our "mixed" model (a weighted average of the SNAIVE, ETS, and other models we used) seemed to give a lower RMSE.
❖ RMSE For Random Forest Model (Figure 3)
  ❖ The RMSE error score is higher for data sets that have fewer observations.
    ❖ Reason: less data to use for training data, so the predictions are not as accurate when run on test data
    ❖ Note, even with a high count, some data is much more sporadic and so harder to predict
❖ RMSE For XGBoost Model (Figure 4)
  ❖ Results were promising, but still had a high RMSE value due to the small dataset.
  ❖ Accounted for potential overfitting and underfitting using cross validation.

FUTURE GOALS

Current Constraints:
❖ Data Familiarity: Since we had a limited amount of time, we familiarized ourselves with the internal time and billing database
❖ Technical Experience With Modeling: This was everyone in our team's first time learning to use tidymodels, modeling techniques, and tidyverse. Hence, we had to orient ourselves with the coding style in the beginning.
❖ Data Formatting: We had to collect the needed data and reformat it to be compatible with the modeling process.

Future Focus:
❖ Increase Data Set: Increase the number of geographies, service groups, and industry groups to make the training data frame larger and have the model be more accurate
❖ Reliable, Yearly Forecasting Engine: Combine the work we have done in both semesters to create a model that uses external and internal factors of CLA information and employee breakdown to create an engine used in yearly planning and for leadership
❖ Complete The R Package: Create additional functionality and methods that also run and help leadership when the standard R tasks are run for CLA systems

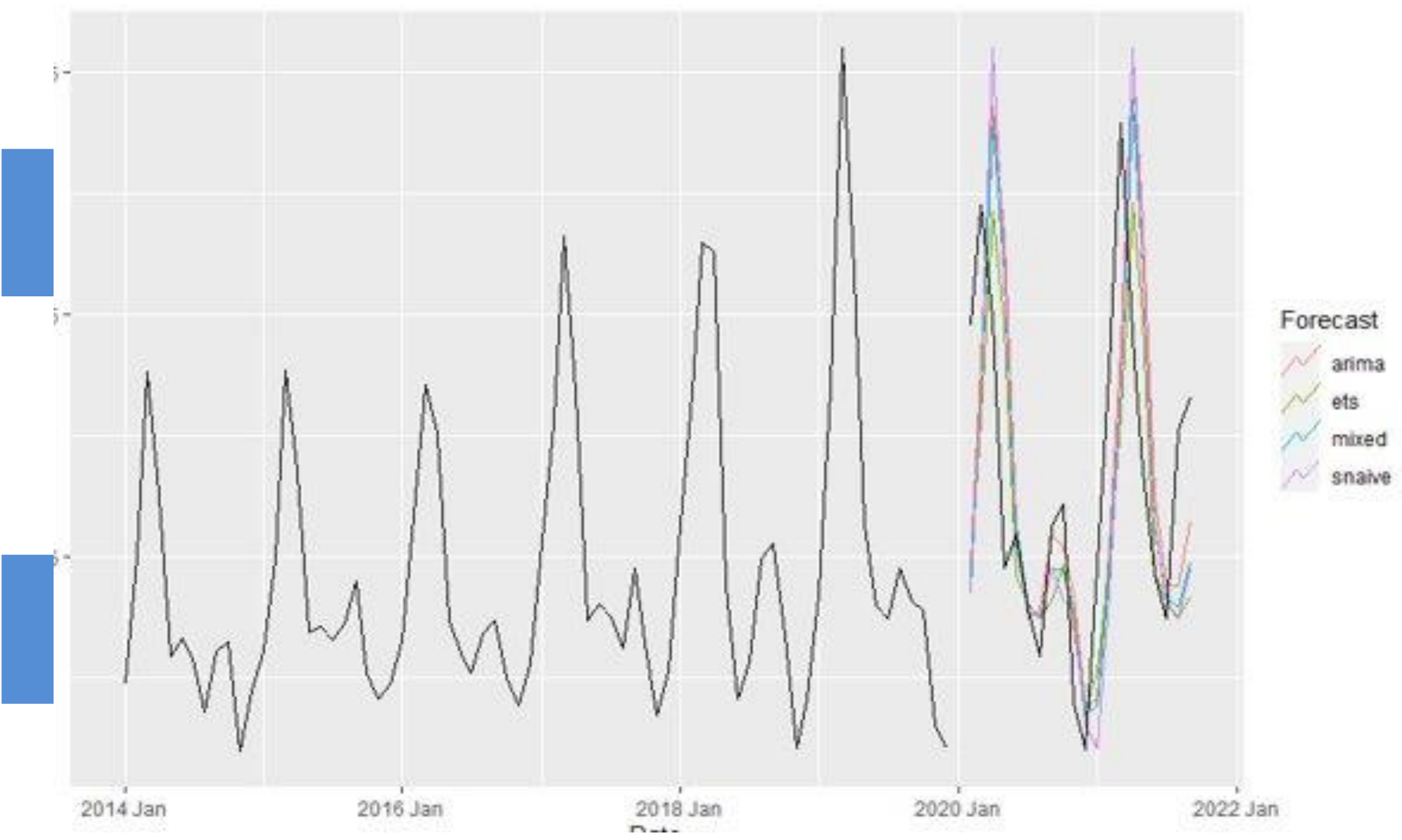

Figure 2: Predictions From Mixed Model

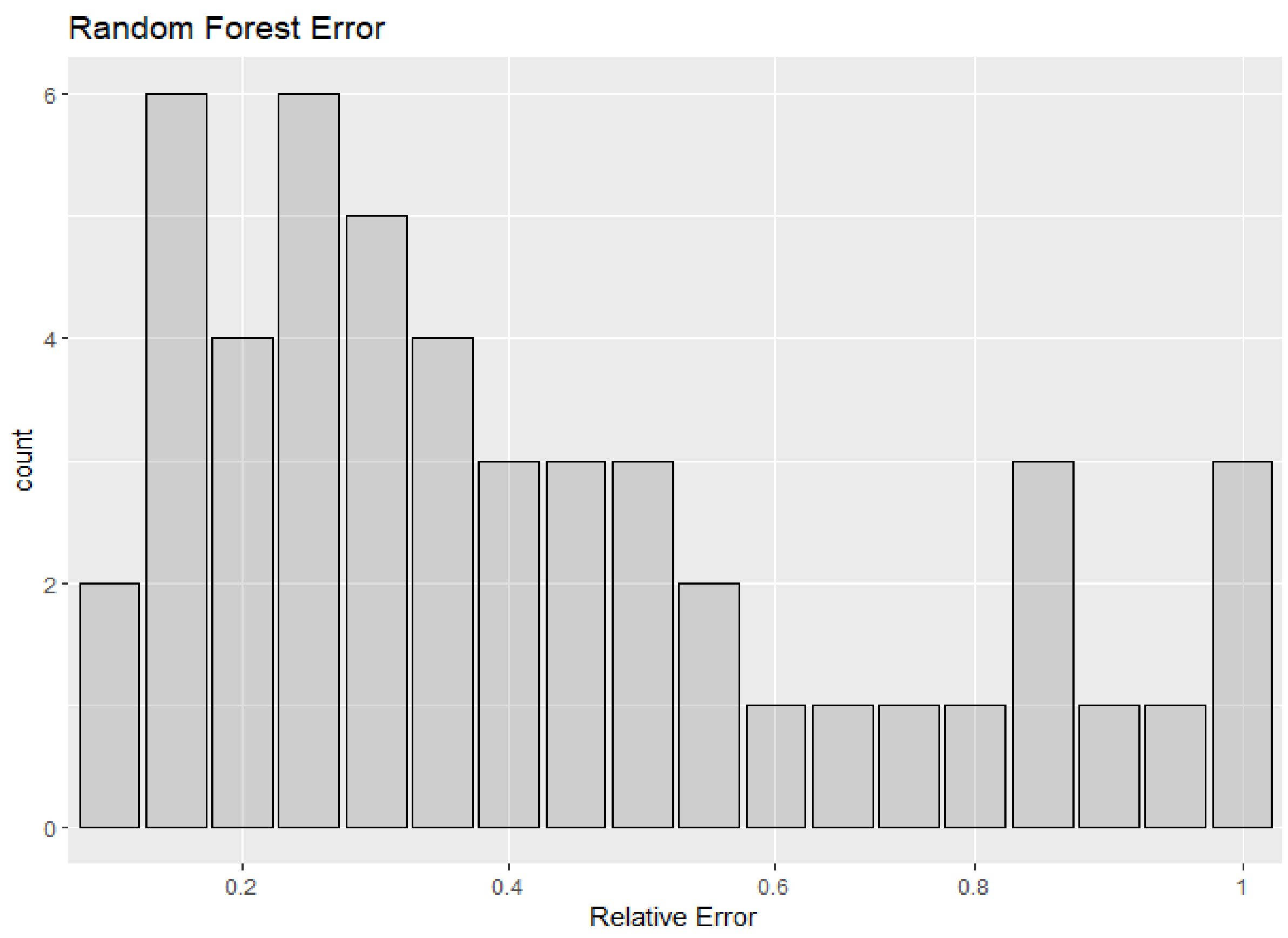

Figure 3: Residual Error Rate From Random Forest Modeling

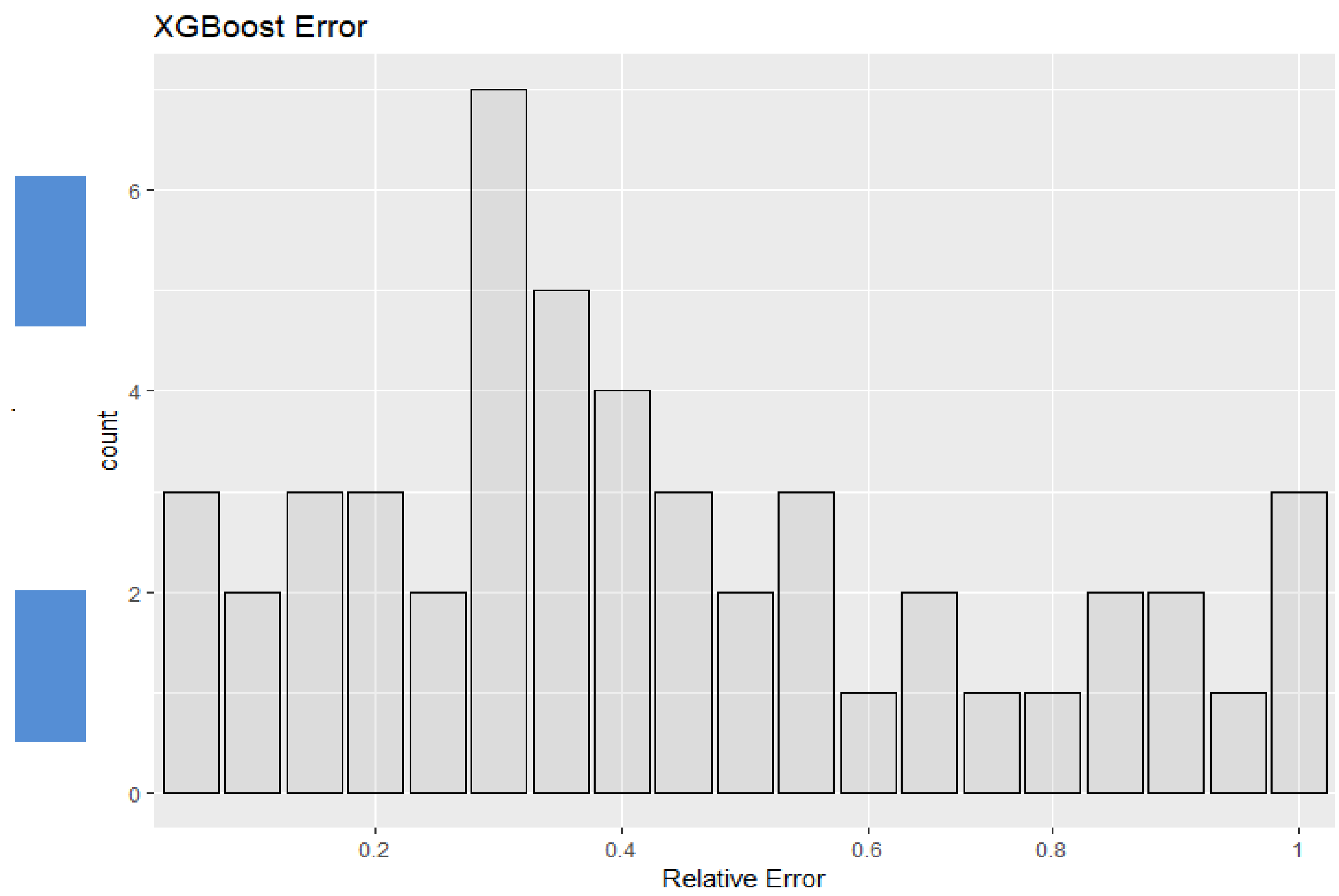

Figure 4: Residual Error Rate From XGBoost Modeling

ACKNOWLEDGEMENTS

Our team would like to profusely thank Dr. Spencer Lourens and Ms. Demi Johnson for being excellent sources of guidance and help throughout this project, helping explain both the technical code and the internal structure of CLA. We would also like to thank the Data Mine Staff, especially Shuennhau Chang, for ensuring our project was running smoothly. Thank you!

* None of the data on this poster is actual CLA data, due to security reasons. *

The Data Mine Corporate Partners Symposium 2022