SESUG Paper 194 - 2023

Predicting Employees’ Preference for Remote or On-site Jobs

Livingstone Gadzanku, Kennesaw State University; Benjamin Watson, Kennesaw State University

ABSTRACT

The rise of remote work has transformed the employment landscape, necessitating effective workforce management strategies. This study utilizes SAS (Statistical Analysis System) to develop a predictive model for employees' remote or on-site work preferences. Leveraging a dataset that encompasses employee demographics, job characteristics, and work-life balance indicators, various SAS analytical techniques, including data exploration, feature engineering, and machine learning algorithms, are employed.

Comprehensive exploratory data analysis and preprocessing are conducted, followed by the utilization of suitable machine learning methods to construct the model. This ensures robust evaluation and addresses class imbalance. The project aims to provide organizations with valuable insights to design optimal remote work policies, enhancing employee satisfaction and productivity.

INTRODUCTION

In the evolving landscape of employment, work preferences are undergoing a profound transformation. The rise of remote work has not only altered the way work is conducted but has also presented new challenges and opportunities for both employees and organizations. As professionals representing diverse industries and roles, the critical importance of adapting to this paradigm shift is apparent.

This paper seeks to explore a fundamental question: How can organizations accurately predict and harness employees' preferences for remote or on-site work? To address this inquiry, the focus is on the realm of predictive modeling, leveraging SAS (Statistical Analysis System). This endeavor offers valuable insights and equips professionals spanning a wide spectrum of roles and industries with the requisite skills and knowledge.

The journey begins with a comprehensive examination of the factors influencing work preferences. Whether one specializes in HR, striving to optimize workforce management strategies, or is a data scientist interested in advanced analytics, this paper furnishes the necessary tools and knowledge to navigate the intricate interplay of employee demographics, job characteristics, and work-life balance indicators.

The skills and expertise to be acquired extend beyond theoretical understanding. Practical experience in SAS analytical techniques, encompassing data exploration, feature engineering, and machine learning algorithms, is offered. This knowledge empowers individuals to construct robust predictive models, capable of providing their organizations with the requisite insights for informed decision-making.

Additionally, it is acknowledged that the challenges posed by imbalanced datasets are a present-day reality in the data-driven world. Thus, guidance is provided on techniques to address class imbalance, ensuring the accuracy and fairness of predictive models.

Ultimately, the skills and knowledge gleaned from this paper hold tangible benefits for professionals across various industries. By comprehending and predicting employees' preferences for remote or on-site work, organizations can tailor their policies, enhancing employee satisfaction, productivity, and overall success.

Embark on this journey to explore the future of work preferences, where the acquisition of skills and insights is essential for thriving in an ever-changing employment landscape.

FACTORS INFLUENCING WORK PREFERENCES

In this section, this paper explores the multifaceted factors that play a pivotal role in shaping employees' work preferences. These factors encompass a range of dimensions, including:

EMPLOYEE DEMOGRAPHICS

- **Age:** Analyzing how age influences the inclination towards remote or on-site work. Are younger employees more inclined to work remotely, or does the trend vary across different age groups?
- **Gender:** Investigating whether there are gender-based disparities in work preferences and how organizations can address these differences.
- **Time Spent Remote Working Before Covid**: This variable captures the extent of remote work experienced by employees in 2019 (before Covid).

JOB CHARACTERISTICS

- **Industry:** The industry in which an employee works can significantly impact their work preferences. Different industries have distinct operational requirements, cultures, and norms that influence whether remote or on-site work is more practical or preferred. Here, we explore the role of industry in shaping these preferences.
- **Job Role (Managerial vs. Non-Managerial):** The distinction between managerial and non-managerial roles within an organization is another crucial dimension in understanding work preferences. Employees in managerial positions often have different responsibilities and requirements compared to non-managerial staff.
- **Pandemic Effects:** The COVID-19 pandemic accelerated remote work adoption across industries. Exploring how industries adapted to remote work during the pandemic and whether these changes are likely to persist post-pandemic is a critical consideration.
- **Perceptions of Remote Work Before Covid**: Employees' perceptions regarding remote work in before Covid encompass several statements:

  a. "My organization encouraged people to work remotely."

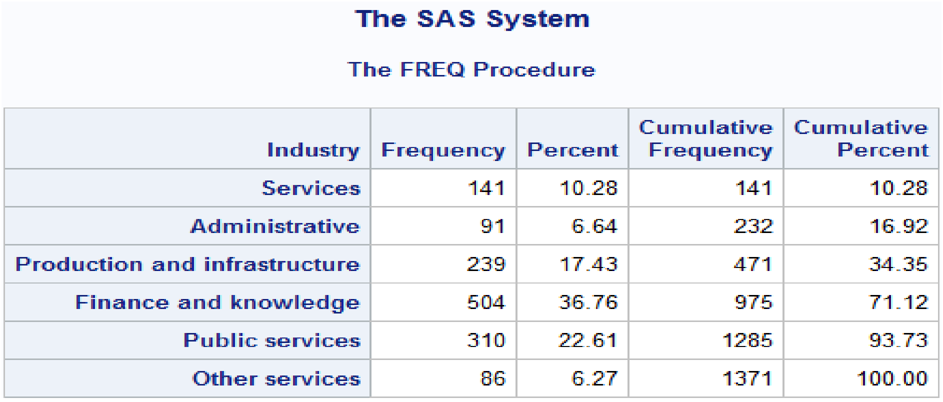
  b. "My organization was well prepared for me to work remotely."

  c. "It was common for people in my organization to work remotely."

  d. "It was easy to get permission to work remotely."

  e. "I could easily collaborate with colleagues when working remotely."

  f. "I would recommend remote working to others."
- **Preferred Time for Remote Work Before Covid**: This variable reflects an individual's preferred work arrangement, aligning with the Job Characteristics category as it relates to an employee's desired work setup.



The SAS System

The FREQ Procedure

Industry	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Services	141	10.28	141	10.28
Administrative	91	6.64	232	16.92
Production and infrastructure	239	17.43	471	34.35
Finance and knowledge	504	36.76	975	71.12
Public services	310	22.61	1285	93.73
Other services	86	6.27	1371	100.00

Table 1. Gender: Number of Males/Females

WORK-LIFE BALANCE INDICATORS

- **Family Status:** Examining how family status, including marital status and the presence of children, impacts work preferences. Are employees with families more likely to opt for remote work to achieve a better work-life balance?
- **Commute Time:** Assessing the relationship between commute time and work preferences. Do employees with longer commutes tend to prefer remote work to reduce commuting stress?
- **Time Spent Remote Working During Covid**: This variable signifies the work arrangements that employees have experienced during covid, significantly influencing work-life balance.
- **Perceptions of Remote Work During Covid**: Similar to the perceptions regarding remote work before covid, this category includes statements assessing recent experiences:

    a. "My organization encouraged people to work remotely."

    b. "My organization was well prepared for me to work remotely."

    c. "It was common for people in my organization to work remotely."

    d. "It was easy to get permission to work remotely."

    e. "I could easily collaborate with colleagues when working remotely."

    f. "I would recommend remote working to others."

- **Preferred Time for Remote Work During Covid:** This variable gauges individuals' preferences for recent work arrangements, making it part of the Work-Life Balance Indicators category.

PRODUCTIVITY

- **Productivity Comparison When Working Remotely**: This question delves into productivity, comparing an individual's work output and quality when working remotely to that at their employer's workplace. It falls under the Productivity category.

These categorized variables constitute the foundation for the comprehensive analysis of the factors influencing work preferences. In the subsequent sections of this paper, we will employ analytical techniques to unravel the intricate relationships between these variables and employees' work preferences. This holistic approach will enable organizations to make informed decisions and tailor their policies effectively in response to the evolving landscape of work preferences.

DATA SOURCE AND METHODOLOGY

In this section, the data source, data preprocessing steps, and methodology applied in constructing the predictive model for employees' remote or on-site work preferences are presented.

DATA SOURCE

The primary source of data for this study was Kaggle, accessible at www.kaggle.com. Kaggle is a renowned repository of datasets and a central hub for data science competitions, serving as a valuable resource for researchers and data analysts in search of real-world data.

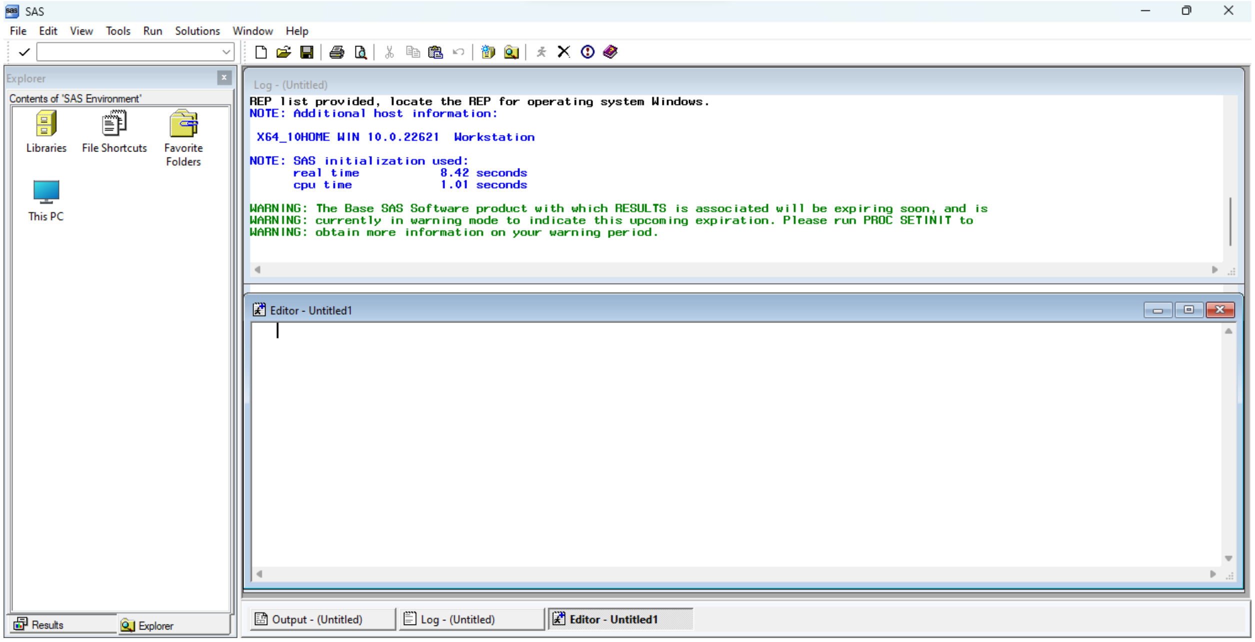
ORIGINAL DATASET

The original dataset procured from Kaggle comprised 1509 observations and featured a total of 73 variables. These variables collectively provided a comprehensive perspective on a range of factors that potentially influence employees' work preferences. Nevertheless, the pursuit of analytical efficiency and the need to simplify the dataset prompted the execution of several data preprocessing measures:

- **Variable Reduction**: Aiming to streamline the analysis and enhance interpretability, a process of variable reduction was executed. This process involved eliminating redundant and excessively detailed responses, including highly specific industry descriptions.
- **Recoding Responses to Numeric Values**: To facilitate the application of machine learning algorithms, all categorical responses within the dataset were systematically transformed into numeric values. This conversion rendered the categorical data amenable to integration into the predictive modeling process.
- **Handling Non-Responses**: It is recognized that the presence of missing or non-responses within a dataset can introduce bias and inaccuracies into the analytical outcomes. In this study, a total of 136 observations, constituting approximately 9% of the overall dataset, were identified as non-responses in relation to the outcome variable concerning work preferences. To preserve data integrity and ensure statistical robustness, these non-responses were excluded from the dataset.

The described data preprocessing measures were pivotal in establishing a dataset of high quality and appropriateness for predictive modeling purposes. The resultant dataset, featuring 35 pertinent variables and numeric responses, served as the foundational material for crafting the predictive model employing SAS analytical techniques, as elaborated upon in subsequent sections of this paper.

Through the meticulous preparation of data and the resolution of non-response issues, the objective was to construct a dataset capable of generating precise and actionable insights into employees' work preferences, ultimately providing organizations with invaluable guidance for their workforce management strategies.



```
REP list provided, locate the REP for operating system Windows.
NOTE: Additional host information:

 X64_10HOME WIN 10.0.22621  Workstation

NOTE: SAS initialization used:
      real time           8.42 seconds
      cpu time            1.01 seconds

WARNING: The Base SAS Software product with which RESULTS is associated will be expiring soon, and is
WARNING: currently in warning mode to indicate this upcoming expiration. Please run PROC SETINIT to
WARNING: obtain more information on your warning period.
```

Display 1. SAS Interface

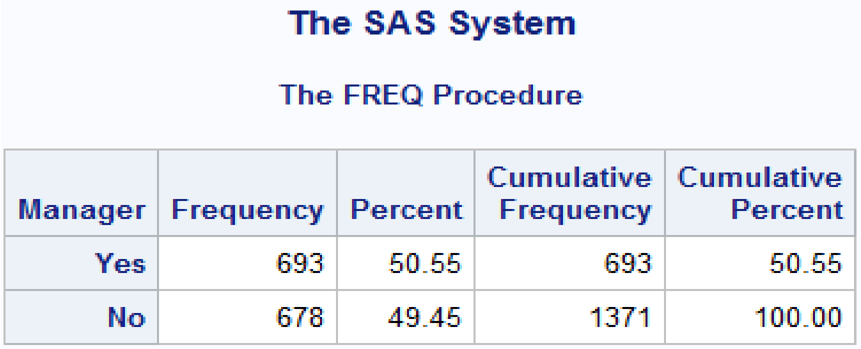
EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics and trends within the dataset. This preliminary analysis lays the foundation for the predictive modeling efforts.

Descriptive Statistics

To gain an initial understanding of the dataset, key descriptive statistics and frequency distributions for relevant variables are presented. Here are summaries for two important variables:

- **Manager Frequency Distribution**: The "Manager" variable provides insights into the distribution of managerial roles among the employees in the dataset:



The SAS System

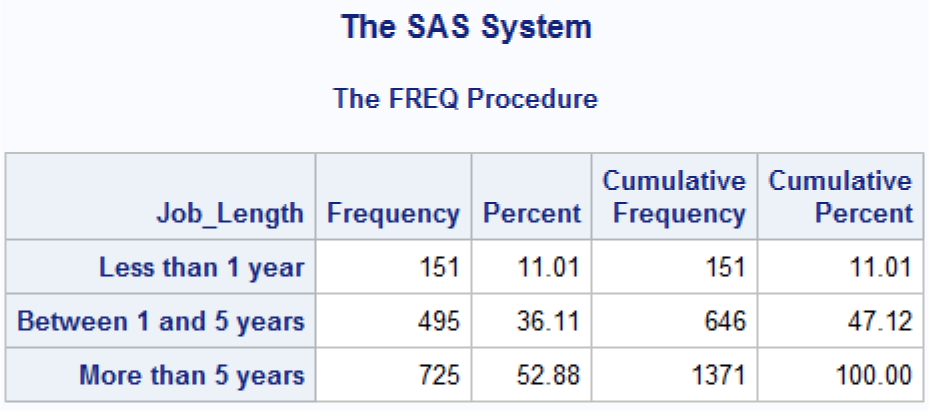
The FREQ Procedure

Manager	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Yes	693	50.55	693	50.55
No	678	49.45	1371	100.00

Table 1. Role: Managerial/Non-Managerial.

This distribution indicates that approximately 50.55% of the employees in the dataset hold managerial positions, while 49.45% do not.

- **Job Length Frequency Distribution**: The "Job_Length" variable provides insights into the distribution of job tenure among the employees:



The SAS System

The FREQ Procedure

Job_Length	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Less than 1 year	151	11.01	151	11.01
Between 1 and 5 years	495	36.11	646	47.12
More than 5 years	725	52.88	1371	100.00

Table 2. Showing Distribution of Job Tenure

This distribution highlights the tenure diversity among employees, with 11.01% having less than 1 year of job experience, 36.11% having between 1 and 5 years, and 52.88% having more than 5 years of experience.

Correlation Analysis

In this section, the results of the correlation analysis are presented, focusing on the relationships between various variables in the dataset. The analysis was conducted using the SAS software, with the computation of Pearson correlation coefficients. The dataset comprises 35 variables, encompassing both outcome and predictor variables.

Pearson correlation coefficients were computed to explore the relationships between these variables. The correlation matrix indicates both the strength and direction of these relationships, with an assessment of significance levels (p-values) to determine statistical significance.

Some findings from the correlation analysis include:

- Outcome and Covid_Remote_Collaborate: A positive correlation of 0.21861 ($p < 0.0001$) is observed between the outcome variable and Covid_Remote_Collaborate. This suggests that increased collaboration during remote work is associated with higher outcome scores.
- Outcome and Covid_Remote_Hrs: A positive correlation of 0.41746 ($p < 0.0001$) is noted between the outcome variable and Covid_Remote_Hrs. This indicates that more hours spent on remote work are associated with higher outcome scores.

- Outcome and Manager: A negative correlation of -0.09833 (p = 0.0003) exists between the outcome variable and the manager variable, suggesting that the presence of a manager is associated with lower outcome scores.

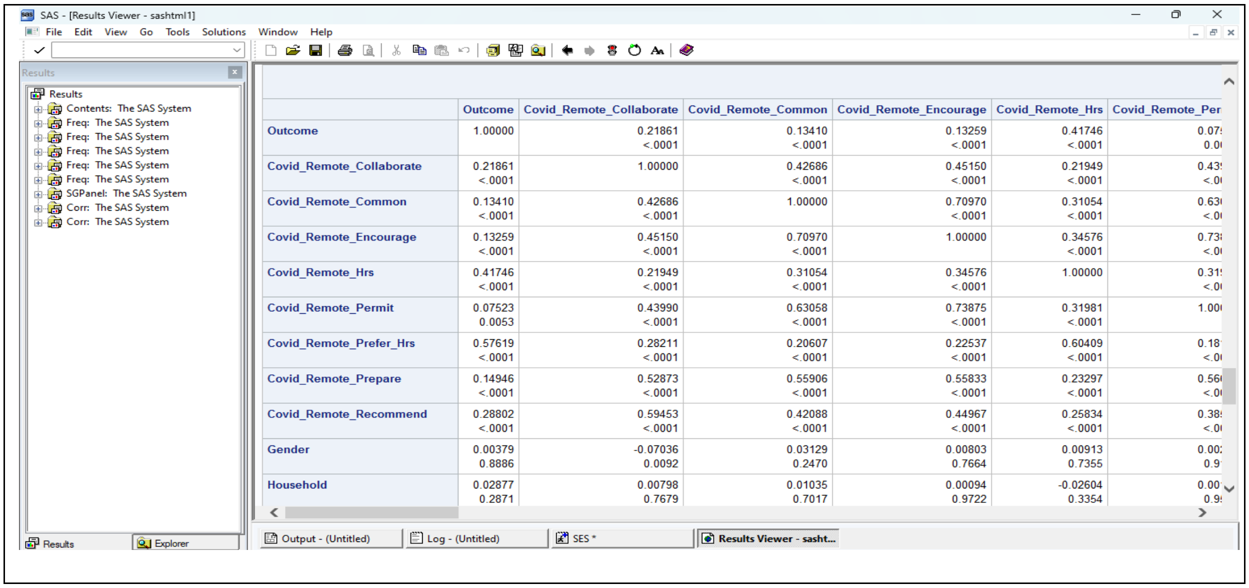
Interpretation:

The correlation analysis provides valuable insights into the relationships between various factors and the outcome variable. These findings offer guidance for further investigation and help in gaining a deeper understanding of the factors influencing outcomes in remote work scenarios. For instance, the positive correlation between collaboration and outcome implies that promoting collaborative remote work practices may lead to better outcomes.

It is important to emphasize that correlation does not imply causation. While associations between variables are observed, further research is required to establish causal relationships, thus providing a foundation for informed decision-making in remote work environments.

Here is the SAS code used to format some variables:

```
/* Format */
Proc format;
      Value Preference
            1 = "Remote"
            2 = "Onsite"
            0 = "NA";

      Value Sex
            1 = "Male"
            2 = "Female"
            0 = "NA";

      Value Industry
            1 = "Services"
            2 = "Administrative"
            3 = "Production and infrastructure"
            4 = "Finance and knowledge"
            5 = "Public services"
            6 = "Other services";

      Value Manager
            1 = "Yes"
            2 = "No"
            0 = "NA";

      Value Household
            1 = "With dependent children"
            2 = "Without dependent children"
            3 = "Single";

      Value Job_Length
            1 = "Less than 1 year"
            2 = "Between 1 and 5 years"
            3 = "More than 5 years";

      Value Rating
            0 = "NA"
            1 = "Strongly Agree"
            2 = "Somewhat Agree"
            3 = "Neither Agree nor Disagree"
            4 = "Somewhat Disagree"
            5 = "Strongly Disagree";

      Value Productivity
            1 = "More productive working remotely"
            2 = "Less productive working remotely"
            0 = "About the same";
Run;
```



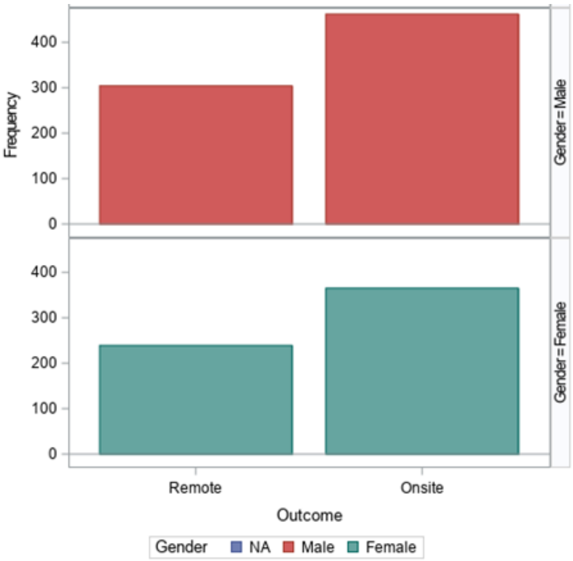
Results
- Contents: The SAS System
- Freq: The SAS System
- Freq: The SAS System
- Freq: The SAS System
- Freq: The SAS System
- Freq: The SAS System
- Freq: The SAS System
- SGPanel: The SAS System
- Corr: The SAS System
- Corr: The SAS System

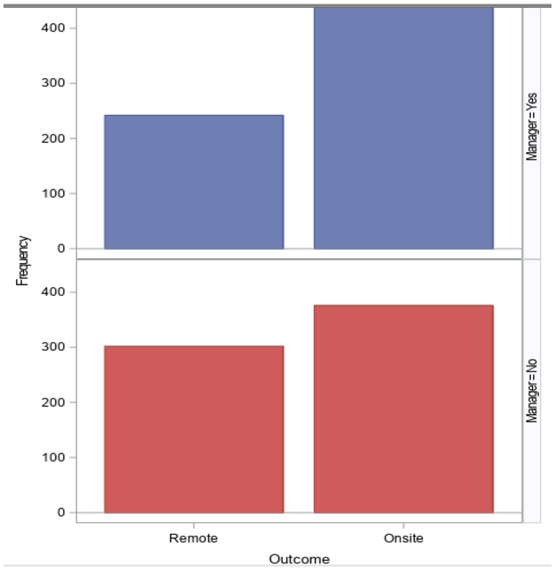
	Outcome	Covid_Remote_Collaborate	Covid_Remote_Common	Covid_Remote_Encourage	Covid_Remote_Hrs	Covid_Remote_Per
Outcome	1.00000	0.21861 <.0001	0.13410 <.0001	0.13259 <.0001	0.41746 <.0001	0.07 0.0
Covid_Remote_Collaborate	0.21861 <.0001	1.00000	0.42686 <.0001	0.45150 <.0001	0.21949 <.0001	0.43 <.0
Covid_Remote_Common	0.13410 <.0001	0.42686 <.0001	1.00000	0.70970 <.0001	0.31054 <.0001	0.63 <.0
Covid_Remote_Encourage	0.13259 <.0001	0.45150 <.0001	0.70970 <.0001	1.00000	0.34576 <.0001	0.73 <.0
Covid_Remote_Hrs	0.41746 <.0001	0.21949 <.0001	0.31054 <.0001	0.34576 <.0001	1.00000	0.31 <.0
Covid_Remote_Permit	0.07523 0.0053	0.43990 <.0001	0.63058 <.0001	0.73875 <.0001	0.31981 <.0001	1.00
Covid_Remote_Prefer_Hrs	0.57619 <.0001	0.28211 <.0001	0.20607 <.0001	0.22537 <.0001	0.60409 <.0001	0.18 <.0
Covid_Remote_Prepare	0.14946 <.0001	0.52873 <.0001	0.55906 <.0001	0.55833 <.0001	0.23297 <.0001	0.56 <.0
Covid_Remote_Recommend	0.28802 <.0001	0.59453 <.0001	0.42088 <.0001	0.44967 <.0001	0.25834 <.0001	0.38 <.0
Gender	0.00379 0.8886	-0.07036 0.0092	0.03129 0.2470	0.00803 0.7664	0.00913 0.7355	0.00 0.9
Household	0.02877 0.2871	0.00798 0.7679	0.01035 0.7017	0.00094 0.9722	-0.02604 0.3354	0.00 0.9

Output 1. Output from a PROC CORR Statement

Preliminary Insights

This stage presents some preliminary insights drawn from the EDA, such as the relationship between managerial roles and work preferences and the distribution of work preferences based on job tenure. These insights offer a glimpse into the dataset's characteristics and set the stage for further exploration and the development of the predictive model.





Output 2. Gender Preference

Output 3. Managerial/Non-managerial preference

FEATURE ENGINEERING

In the context of developing a predictive model, feature engineering assumes a pivotal role, serving to enhance the model's performance by facilitating the selection, transformation, or creation of new features from the dataset. This process aims to improve the model's predictive accuracy and interpretability.

In the analytical pursuit, a simpler model was constructed initially, one encompassing a subset of features. The central consideration throughout this endeavor was the delicate equilibrium to be struck between model simplicity and predictive performance.

Simpler Model (Step 6):

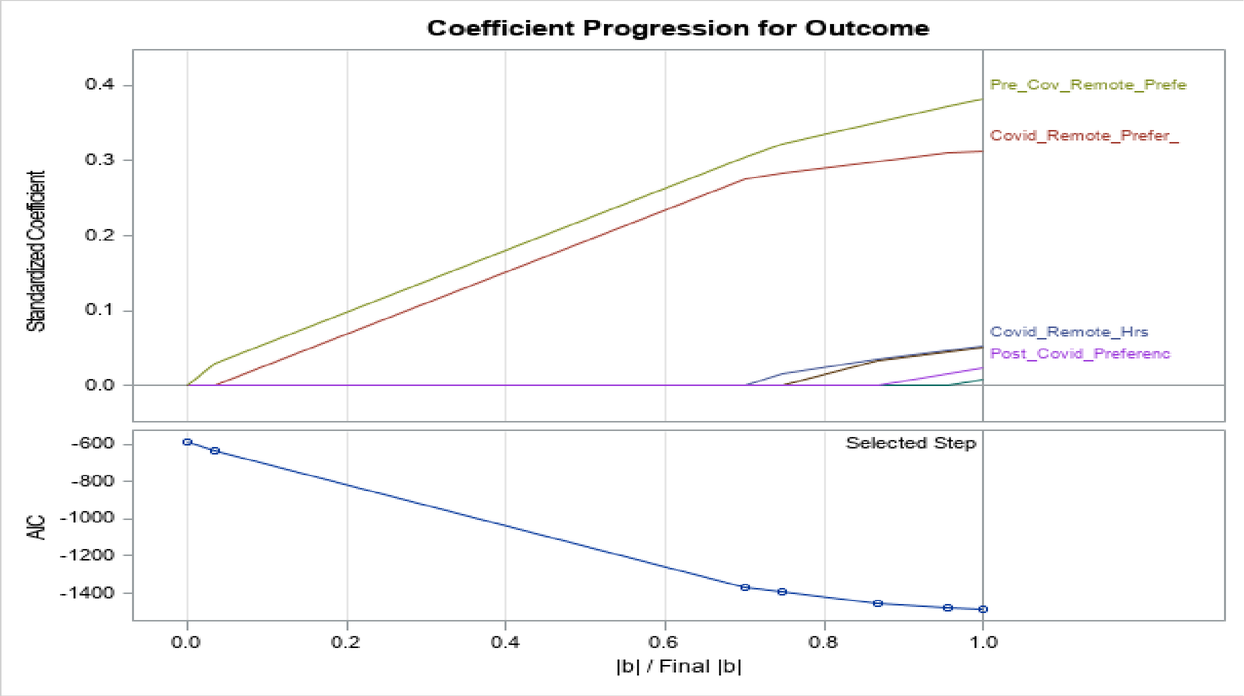
The simpler model, identified at Step 6, was characterized by a more frugal selection of seven distinct features. This assembly bestowed a relatively more concise representation of the underlying data. While the simplicity inherent in this model possesses its own appeal, the pivotal evaluation lies in its performance vis-à-vis an alternative, albeit more intricate model (Step 15) encompassing a grander set of 16 features.

Significant Trade-off:

Of particular note in the analysis is the remarkable balance achieved by the simpler model, despite its more conservative feature set, between model complexity (as gauged by AIC) and predictive accuracy (as assessed by RMSE), in comparison to the more intricate counterpart.

In simpler terms, the trade-off in contemplation is this: the simpler model, distinguished by its reduced feature count, exhibits a performance that closely shadows the more elaborate model (boasting 16 variables) regarding predictive accuracy (RMSE) and model fit (AIC). This underscores the notion that the simpler model does not significantly lag behind the complex model in its ability to elucidate the variance in the response variable, all while preserving a more interpretable and manageable model structure.

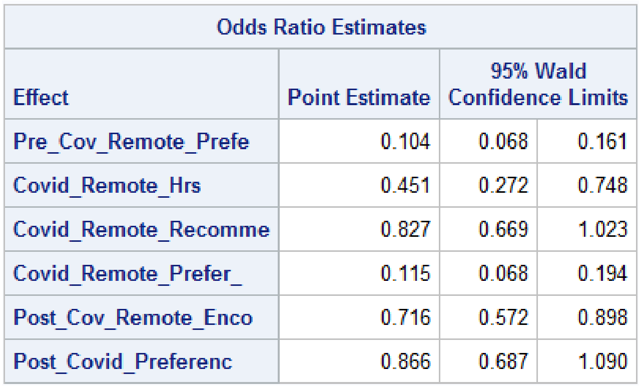
In light of these considerations, the selection between these models hinges upon the precise objectives of the analysis. If simplicity, interpretability, or computational efficiency take precedence, the simpler model is poised for preference. Conversely, if the paramount goal revolves around optimizing predictive performance, with the accompanying increase in complexity deemed warranted, the more intricate model may serve as the suitable choice. The ultimate decision should be guided by the trade-off that most harmoniously aligns with the research objectives and pragmatic constraints.



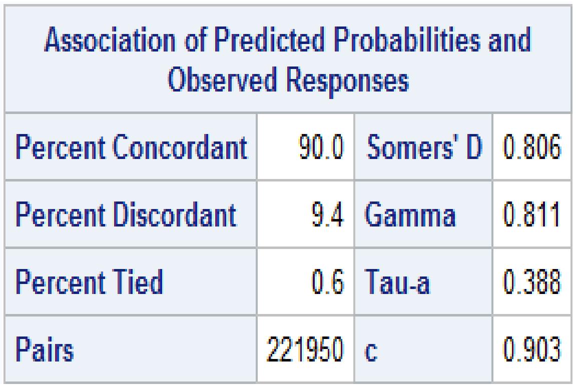
Output 4. Output from GLM SELECT PROCEDURE

MODEL EVALUATION

The logistic regression model's performance evaluation reveals several critical insights, emphasizing its capability to predict employees' job preferences effectively. Key elements extracted from the evaluation tables are presented below:



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Pre_Cov_Remote_Prefe	0.104	0.068	0.161
Covid_Remote_Hrs	0.451	0.272	0.748
Covid_Remote_Recomme	0.827	0.669	1.023
Covid_Remote_Prefer_	0.115	0.068	0.194
Post_Cov_Remote_Enco	0.716	0.572	0.898
Post_Covid_Preferenc	0.866	0.687	1.090



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	90.0	Somers' D	0.806
Percent Discordant	9.4	Gamma	0.811
Percent Tied	0.6	Tau-a	0.388
Pairs	221950	c	0.903

Table 3. Odds Ratio Estimates **Table 4. Assoc. of Pred. Prob. And Obs. Responses**

The Odds Ratio Estimates signify the impact of predictor variables on the likelihood of a particular job preference outcome. These estimates, along with their respective 95% Wald Confidence Limits, allow us to gauge the significance and direction of influence for each predictor.

Association of Predicted Probabilities and Observed Responses provides essential metrics that assess the alignment between predicted probabilities and observed responses. These metrics include:

- Percent Concordant: Reflects the percentage of pairs where predicted probabilities and actual outcomes are correctly ordered.
- Somers' D: Measures the strength and direction of association between predicted probabilities and observed outcomes.
- Percent Discordant: Represents the percentage of pairs with discordant predictions and observed outcomes.
- Gamma: Indicates the degree of agreement between predicted and observed outcomes.
- Percent Tied: Accounts for tied predictions in pairs.
- Tau-a: Measures the degree of association between predicted probabilities and observed outcomes, considering tied values.
- Pairs: The total number of pairs considered in the analysis.
- c: Reflects the model's overall predictive ability, with values closer to 1 indicating better performance.

These key elements collectively attest to the model's capacity to accurately predict job preferences and provide valuable insights for organizational decision-making.

CONCLUSION

In summary, this study investigated employees' remote and onsite work preferences through predictive modeling. Two models, one complex and one simplified, demonstrated the trade-off between model complexity and interpretability. Model evaluation highlighted the balance between AIC and RMSE metrics. Cross-validation reinforced the model's reliability.

In conclusion, this research offers practical insights for organizations seeking to align work arrangements with employee preferences, providing a valuable predictive model for shaping the future of work

ACKNOWLEDGMENTS

We would like to express our gratitude to Professor Joe DeMaio for introducing us to this opportunity and for providing invaluable guidance throughout this research endeavor.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Livingstone Gadzanku
Phone: 4709204894
E-mail: eligadzanku@gmail.com or lgadzank@students.kennesaw.edu

Name: Benjamin Watson
Phone: 7706057974
E-mail: benjaminwatson24@gmail.com or bwatso64@students.kennesaw.edu