**Human Resources Analytics at CenturyLink**

Kyle Carey

Colorado State University – Global Campus

MIS581: Capstone in Data Analytics

Dr. Osama Morad

8/23/2020

**Description of Dataset**

The dataset I chose for this assignment is a human resources dataset published by Dr. Rich Huebner describing employees in a fictitious organization (Dr. Huebner, 2019). Due to the inherent sensitivity of employee data, I chose to use an external dataset to act as a real-world example of data at CenturyLink; while the data for this project is synthesized, I could easily replicate each of the fields in it to perform analyses for CenturyLink. The dataset contains 401 observations of 35 variables, including outcome data such as employee engagement, performance scores, and turnover.

**Tools for Analysis**

To begin data analysis, the dataset needs to be formatted appropriately. Each variable should be coded correctly for analysis, and derived fields, such as age (based on Date of Birth), should be calculated. Any missing values should be addressed, either through removal or calculation. Once the dataset is ready, it needs to be explored. Exploratory analysis provides insight into predictive or prescriptive analyses appropriate for the dataset; for example, a basic correlation matrix across different variables may show trends worthy of further exploration through, say, regression. Next, a variety of analytical models need to be deployed on the dataset based on research questions and exploratory findings to maximize the predictive power of said models. Similarly, an ensemble model may be deployed to take advantage of the features of different models. Finally, findings throughout the process need to be visualized to communicate results to appropriate stakeholders.

The data analysis process – ETL, variable derivation, preliminary analysis, model deployment, and visualization – will all be handled in R. R is an open-source data analysis software popular for its versatility as both a command-line and scripting language, as well as its diverse and striking visualization packages. Through R, two main tools (packages) will be leveraged:

1. Tidyverse, a package ensembled designed to simply and streamline data management, and
2. ggplot2, a package that delivers an object-oriented approach to generating graphics.

Data Dictionary

Below is a data dictionary for the datatset:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **DataType** |
| Employee Name | Employee’s full name | Text |
| EmpID | Employee ID is unique to each employee | Text |
| MarriedID | Is the person married (1 or 0 for yes or no) | Binary |
| MaritalStatusID | Marital status code that matches the text field MaritalDesc | Integer |
| EmpStatusID | Employment status code that matches text field EmploymentStatus | Integer |
| DeptID | Department ID code that matches the department the employee works in | Integer |
| PerfScoreID | Performance Score code that matches the employee’s most recent performance score | Integer |
| FromDiversityJobFairID | Was the employee sourced from the Diversity job fair? 1 or 0 for yes or no | Binary |
| PayRate | The person’s hourly pay rate. All salaries are converted to hourly pay rate | Float |
| Termd | Has this employee been terminated - 1 or 0 | Binary |
| PositionID | An integer indicating the person’s position | Integer |
| Position | The text name/title of the position the person has | Text |
| State | The state that the person lives in | Text |
| Zip | The zip code for the employee | Text |
| DOB | Date of Birth for the employee | Date |
| Sex | Sex - M or F | Text |
| MaritalDesc | The marital status of the person (divorced, single, widowed, separated, etc) | Text |
| CitizenDesc | Label for whether the person is a Citizen or Eligible NonCitizen | Text |
| HispanicLatino | Yes or No field for whether the employee is Hispanic/Latino | Text |
| RaceDesc | Description/text of the race the person identifies with | Text |
| DateofHire | Date the person was hired | Date |
| DateofTermination | Date the person was terminated, only populated if, in fact, Termd = 1 | Date |
| TermReason | A text reason / description for why the person was terminated | Text |
| EmploymentStatus | A description/category of the person’s employment status. Anyone currently working full time = Active | Text |
| Department | Name of the department that the person works in | Text |
| ManagerName | The name of the person’s immediate manager | Text |
| ManagerID | A unique identifier for each manager. | Integer |
| RecruitmentSource | The name of the recruitment source where the employee was recruited from | Text |
| PerformanceScore | Performance Score text/category (Fully Meets, Partially Meets, PIP, Exceeds) | Text |
| EngagementSurvey | Results from the last engagement survey, managed by our external partner | Float |
| EmpSatisfaction | A basic satisfaction score between 1 and 5, as reported on a recent employee satisfaction survey | Integer |
| SpecialProjectsCount | The number of special projects that the employee worked on during the last 6 months | Integer |
| LastPerformanceReviewDate | The most recent date of the person’s last performance review. | Date |
| DaysLateLast30 | The number of times that the employee was late to work during the last 30 days | Integer |

Data Value to CenturyLink

This dataset provides value to an organization such as Centurylink in a variety of ways: first, demographic data in the set allows for group differences analysis across outcomes, potentially identifying discrimination or a lack of diversity within the organization; two, it provides the opportunity to predict higher performers within the organization, through recruitment sourcing, etc.; three, it allows the organization to identify managers whose direct reports have low engagement scores and potentially provide ways to improve engagement. While these examples prove the value of human resources datasets, the final analyses for this project may differ after a preliminary exploration of the dataset.

Plenty of research in the past decade supports the value of diversity within an organization. Gartner predicts that through the next few years, more diverse organizations will exceed their financial targets and outperform less inclusive organizations by as much as 50%. As much as 83% of millennials report a higher level of employee engagement when their employer emphasizes a diverse and inclusive culture. In the technology industry, higher turnover of underrepresented groups accounts for more than $16 billion in costs each year (Purdue University Global, 2020). By identifying underrepresented populations in a dataset, the organization may take action to increase diversity and realize better outcomes because of it.

Turnover costs organizations. In general, according to Li (2018), the cost of turnover reaches 21% of an employee’s annual salary (in marketing, recruitment, and re-training). Because of this, sourcing candidates who have a longer days-until-termination than average could provide a significant return on investment for an organization. The dataset for analysis in the project includes start date, termination date, and termination reason, along with potential mitigating factors such as recruiting source, pay, and manager. This dataset may provide insight into which factors predict longer days-until-termination and allow the organization to predict employees with greater-than-average tenure.

Finally, the dataset for this organization allows for the analysis of pay equity, or the fairness of pay across demographic groups. A variety of US Federal law requires pay equity across an organization, including the Equal Pay Act, the Civil Rights Act, the Age Discrimination in Employment Act, the Americans with Disabilities Act, and the Fair Pay Act of 2007 (SHRM, 2020). (As an aside, a limitation of this dataset is that it does not provide benefit information such as a time off rate, healthcare coverage, leave options, bonuses, or stock options. These and other non-pay benefits may contribute significantly to an employee’s total compensation package.) While salary guidelines need to be set by compensation, human resources analysts can analyze pay rates to identify inequity in pay practices. For this dataset, that may include inequity across gender, race, and/or age.

**Analysis**

Analysis will begin with an approach to understanding and improving diversity within the organization. First, exploratory analyses will help to determine the nature of diversity in the organization. Then, a regression model will attempt to predict a variety of factors contributing to the hiring of a diverse workforce. The next analysis will attempt to predict longer tenure at the organization of those who’ve turned over. Finally, a pay equity analysis, while not predictive in nature, will provide invaluable insights into unconscious bias within the organization.

**Ethics in HR Data Analysis**

Human Resources departments have access to a seemingly unprecedented amount of sensitive employee data, from performance history, demographics information, and background check summaries to ADA accommodation requests, social security numbers, and pay grade information. While data protection laws in the United States such as HIPAA protect a portion of this data, individual Human Resources departments need to create policies and procedures for using data not necessarily protected by law. This report will outline the potential ethical challenges presented by the dataset used for the capstone analysis as well as provide an overview on the policies and procedures this Human Resources department may create to help overcome those challenges.

**Potential Ethical Challenges**

To begin, the dataset for the capstone analysis includes sensitive information attached to employee names and ID numbers, including pay grade, gender, race/ethnicity, and marital status. Because the capstone analyses do not involve individual action, these identifying columns can be dropped and replaced with a surrogate identifier.

In Human Resources predictive analytics, the application of predictive models needs to be monitored for adverse impact. For example, a model that predicts turnover may disparately impact women in the workplace if women have gender-related reasons for exiting the business (such as pregnancy). Because of this, results from a test dataset should be evaluated through a common adverse impact analysis (chi squared, for example). If the predictive model shows no potential for adverse impact, it may be used on candidate or employee data to better improve organizational outcomes.

In general, data collection in the United States should follow California’s Consumer Data Protection Act, including (State of California - Department of Justice - Office of the Attorney General, 2020):

* The right to know what information a business collects and how it’s used;
* The right to have deleted personal information collected or analyzed;
* The right to opt out of the sale of personal information; and
* The right to non-discrimination.

**Business Goal 1: Increasing Diversity**

Plenty of research in the past decade supports the value of diversity within an organization. Gartner predicts that through the next few years, more diverse organizations will exceed their financial targets and outperform less inclusive organizations by as much as 50%. As much as 83% of millennials report a higher level of employee engagement when their employer emphasizes a diverse and inclusive culture. In the technology industry, higher turnover of underrepresented groups accounts for more than $16 billion in costs each year (Purdue University Global, 2020). By identifying underrepresented populations in a dataset, the organization may take action to increase diversity and realize better outcomes because of it.

Business Goal: Increase diversity throughout the organization.

Research Question 1: Which recruitment source funnels the most diverse applicant pools that result in diversified hiring?

H0: No differences exist between the demographics of hires between recruitment sources.

Ha: At least one recruitment source funnels more diverse applicants (resulting in hires) than others.

**Testing Methodology**

Diversification analysis for this research question will begin with the recoding of demographic data to code diversification as a binary variable: 0 for white males, 1 for minority (race/ethnicity and/or gender). After an exploratory analysis of the data, regression will help determine which recruiting sources funnel more diverse hires, if any. Data preparation, analysis, and visualization will be in R.

Research Question 2: Can we predict based on recruitment source, state, and position whether a hire will be “diverse” or not?

H0: Recruitment source, state, and position do not predict diversification.

Ha: Recruitment source, state, and position predict diversification.

**Testing Methodology**

Using the same diversification coding as above, a generalized linear model (logistic) will be trained to analyze the effectiveness of predicting diversification in the organization. The dataset will be partitioned into a training and testing set, and both the coefficients from the regression model and a confusion matrix will be presented as evidence for or against rejecting the null hypothesis.

**Results**

The overall accuracy in predicting minority status through a logistic regression model is close to 42%, which is lower than the “no information rate.” Because of this, this test fails to reject the null hypothesis.

Accuracy : 0.4194

95% CI : (0.3407, 0.5012)

No Information Rate : 0.7355

P-Value [Acc > NIR] : 1

Kappa : -0.0055

Mcnemar's Test P-Value : 1.276e-10

Sensitivity : 0.6585

Specificity : 0.3333

Pos Pred Value : 0.2621

Neg Pred Value : 0.7308

Prevalence : 0.2645

Detection Rate : 0.1742

Detection Prevalence : 0.6645

Balanced Accuracy : 0.4959

'Positive' Class : 0

**Business Goal 2: Reducing Turnover**

Turnover costs organizations. In general, according to Li (2018), the cost of turnover reaches 21% of an employee’s annual salary (in marketing, recruitment, and re-training). Because of this, sourcing candidates who have a longer days-until-termination than average could provide a significant return on investment for an organization. The dataset for analysis in the project includes start date, termination date, and termination reason, along with potential mitigating factors such as recruiting source, pay, and manager. This dataset may provide insight into which factors predict longer days-until-termination and allow the organization to predict employees with greater-than-average tenure.

**Business Goal: Identify potential reasons for turnover.**

Research Question 1: What factors contribute to lower days-to-termination for employees?

H0: No combination of factors predicts fewer days to term.

Ha: At least one variable predicts fewer days to term.

In a regression model to predict termination based on pay rate, a regression model predicted turnover correctly 51% of the time, again lower than the no-information rate. The research again fails to reject the null hypothesis.

Reference

Prediction 0 1

0 34 12

1 64 45

Accuracy : 0.5097

95% CI : (0.4282, 0.5907)

No Information Rate : 0.6323

P-Value [Acc > NIR] : 0.9993

Kappa : 0.1146

Mcnemar's Test P-Value : 4.913e-09

Sensitivity : 0.3469

Specificity : 0.7895

Pos Pred Value : 0.7391

Neg Pred Value : 0.4128

Prevalence : 0.6323

Detection Rate : 0.2194

Detection Prevalence : 0.2968

Balanced Accuracy : 0.5682

'Positive' Class : 0

**Conclusion**

Human behavior is complex and often unpredictable. In Human Resources, the ability to predict outcomes such as turnover proves to be an almost invaluable insight for organizations, resulting in tremendous ROI due to the high costs of turnover. Similarly, diverse organizations provide tremendous value to organizations, and recruiting from sources proven to funnel diverse applicants helps in that endeavor. Despite this research failing to reject the null hypotheses of the research, the findings allow us to explore other variables to predict turnover and diversity within the organization.

Appendix

**Full R Code**

library(caret)

data1 <- read.csv("C:/Users/607138/Downloads/HRDataset\_v13.csv")

data1$minority.status <- ifelse(data1$RaceDesc == "White" & data1$Sex == "M ","0","1")

data1$minority.status <- as.numeric(data1$minority.status)

data1$online.source <- ifelse(data1$RecruitmentSource == "Careerbuilder" | data1$RecruitmentSource == "Company Intranet - Partner" | data1$RecruitmentSource == "Glassdoor" | data1$RecruitmentSource == "Indeed" | data1$RecruitmentSource == "Internet Search" | data1$RecruitmentSource == "Monster.com" | data1$RecruitmentSource == "On-line Web application" | data1$RecruitmentSource == "Pay Per Click" | data1$RecruitmentSource == "Pay Per Click - Google" | data1$RecruitmentSource == "Search Engine - Google Bing Yahoo" | data1$RecruitmentSource == "Social Networks - Facebook Twitter etc" | data1$RecruitmentSource == "Website Banner Ads","1","0")

set.seed(1)

train.index <- sample(c(1:310), 155)

traindata <- data1[train.index,]

testdata <- data1[-train.index,]

testdata$minority.status <- factor(testdata$minority.status)

model1 <- glm(minority.status ~ online.source, data = traindata, family = 'binomial')

testdata$pred <- predict(model1, testdata)

testdata$pred <- cut(testdata$pred, c(-Inf, 1, Inf), labels = c(0,1))

confusionMatrix(testdata$pred, testdata$minority.status)

model2 <- glm(Termd ~ PayRate, data = traindata, family = 'binomial')

testdata$termdpred <- predict(model2, testdata)

testdata$termdpred <- cut(testdata$termdpred, c(-Inf, -.8277, Inf), labels = c(0,1))

testdata$Termd <- factor(testdata$Termd)

confusionMatrix(testdata$termdpred, testdata$Termd)

References

California Consumer Privacy Act (CCPA). (2020, July 20). State of California - Department of Justice - Office of the Attorney General. https://oag.ca.gov/privacy/ccpa

Dr. Huebner. (2019, September 29). HR Codebook (Version 13) [Dataset]. https://rpubs.com/rhuebner/HRCodebook-13

Li, M. (2018). Costly Turnover. Fisher College of Business. https://fisher.osu.edu/blogs/leadreadtoday/blog/costly-turnover

Purdue University Global. (2020). Does Workplace Diversity Actually Impact a Business? Purdue Global. https://www.purdueglobal.edu/blog/careers/how-does-workplace-diversity-affect-business/

SHRM. (2020). Managing Pay Equity. Society for Human Resources Management. https://www.shrm.org/resourcesandtools/tools-and-samples/toolkits/pages/managingpayequity.aspx