mgracer7795 / **dsc-capstone-project-v2-3**

Code | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

dsc-capstone-project-v2-3 Public

forked from learn-co-curriculum/dsc-capstone-project-v2-3

main | Go to file | Add file | Code

Branches | Tags

This branch is 4 commits ahead of learn-co-curriculum:main. | Contribute | Sync fork

mgracer7795 Add files via upload ...		5 minutes ago 14
.canvas	add .canvas	2 years ago
.gitignore	Initial commit	2 years ago
FirstSimpleModel.ipynb	Add files via upload	last week
Human Resources_ A Stu...	Add files via upload	5 minutes ago
Phase5Project.ipynb	Add files via upload	9 minutes ago
README.md	Add files via upload	10 minutes ago

README.md

About

No description, website, or topics provided.

Readme
Activity
0 stars
0 watching
10 forks

Releases

No releases published
Create a new release

Packages

Overview

This project involves the comprehensive task of analyzing HR Analytics data from Kaggle. The data contains 38 columns that describe a wide variety of information about employee patterns that either result in attrition or not. Attrition is high in the workplace so the majority of columns from the dataset are categorized as 0 for resulting in attrition. For the purposes of this study and balancing the data Gender was chosen as a measure of impact as the target variable. If attrition was more balanced it would make sense as a target variable, however, the data was skewed toward attrition.

An unbalanced study of attrition was also implemented as a target variable. Linear regressions were run for the numeric variables with years in current role and monthly income used as independent variables. A multi-class study was also conducted for marital status. The marital status variable has three categories instead of two.

Business Objective

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%

A HR Tech company has hired me to analyze this data in an effort to better understand what factors are impacting Attrition levels. Attrition is high in the modern workplace so it is essential for organizations to gain a stronger understanding of what is driving this trend for their specific organization. I've been tasked with identifying the variables that are influencing Attrition patterns. Certain target variables of interest are Gender, Attrition itself, and Monthly Income for a Linear Regression. Gender is a binary variable from this dataset that can be informative to the breakdown between Males and Females when it comes to Attrition levels. Attrition is also a binary variable that is a strong target variable because all other variables are going to be informative of the direct impact on Attrition levels. Attrition was an unbalanced study, however, because the observations were skewed toward mostly having Attrition. Monthly Income was also used as a target variable for a Linear Regression study to present what numeric dependent variables are statistically significant. The HR Tech Company is looking for metrics that demonstrate a strong relationship between the target variable and selected predictive varibles.

Data Understanding and Analysis

The first step of the analysis was to indentify a target variable. Gender was selected because it is a binary variable that is associated with Attrition in the workplace. It was also able to be balanced. Some of the predictive variables that were used were Age, BusinessTravel,DailyRate,Education,Environment Satisfaction,PerformanceRating, and Relationship Satisfaction. The target variable needed to be One Hot Encoded because it was an object type categorized as Male or Female for Gender. One Hot Encoding converted this to 1 for Male and 0 for Female. Then the data was balanced so the number of observations for both genders was even.

A train test set was then completed and then the data needed to be One Hot Encoded. This was completed for both the train and test datasets.

The baseline model for this project is a Random Forest Classifier. Precision, Recall, Accuracy, and an F1 Score were all generated from this model. The scores varied and are far above the minimum value of 0.5. The precision score is 0.75, the Recall score is 0.94, the accuracy score is 0.81, and the F1 Score is 0.84. Cross Validation Scores were also generated and the mean recall score was the highest at 0.60. A ROC curve was also generated and had an accuracy of 0.96. These are promising results and demonstrates that the Random Forest Classifier is performing well on both positive and negative predictions from this dataset.

Three other additional models were added that were Logistic Regression, Decision Tree Classifier, and Multinomial. Classification reports were generated for these three models. The results were lower than the baseline model. The highest score for the Logistic Regression model was 0.57 for recall. The highhest score for the Decision Tree Classifier was 0.53 for recall. For the Mulitnomial model, the recall score of 0.65 was the highest.

A baseline Random Forest Classifier model was also created for an imbalanced study of Attrition. The scores for the Attrition study were an accuracy of 0.84. Precision had a score of 1.0 for the train data. Recall and the F1 Score were skewed toward one variable because the data is unbalanced. This is a favorable outcome as it is closer to a value of 1 than 0.5. A ROC curve was generated for the train data that had a very promising score of 0.96. The central finding here is that the Random Forest Classifier is performing well on both positive and negative predictions for Attrition and Gender.

Multiclass Classification was also implemented for Marital Status. Marital Status had high precision, recall, and f1-scores for the Single and Married categories. The Single status had the highest scores all above 0.70 with the highest being a precision of 0.99. Married had all metric values above 0.6. The divered status had values below 0.5.

The final study was a Multi-Linear Regression model. For this model a diffent target variable was selected since this is a numeric study. The target variable is the YearsInCurrentRole. The predictive variables that were found to have statistical significance at the 1% level are years in current role and monthly rate for the Mulit-Linear Regression. Another target variable was implemented for Monthly Income. The predictive variables that were found to have statistical signficance at the 1% level for this regression were TotalWorkingYears. The constant was significant at the 1% level.

For a standard Linear Regression for YearsInCurrentRole: YearsAtCompany, YearsSinceLastPromotion, YearsInCurrentRole, TotalWorkingYears, and Monthly Income is sigificant at the 1% level. For a standard Linear Regression for MonthlyIncome: YearsAtCompany, YearsSinceLastPromotion, TotalWorkingYears, and YearsInCurrentRole is significant at the 1% level.

Vizualizations

Vizulations are included for scatter plots, histograms, roc curves, confusion Matrix with more iterations and then higher tolerance added for the Logistic Regression model. Roc curves were generated for both Gender and Attrition studies. Linear Regression models were also generated for MonthlyIncome and YearsInCurrentRole as the target variable. Different Linear Regression Models were generated for dependent varibles that were statistically significant. Many of the dependent variables were not sigificant. However, some were sigificant at the 5 and 1% levels. The r**2 values varied.

Recommendations

There are a few different recommendations based on these findings and the recommendations depend on the target variables of each study. The scores for Gender were high for the Random Forest Regressor so that could be a starting point to understand how Gender is influencing Attrition rates. Marital Status had high scores for the multi-class study for the Married and Single categories. You could narrow down the data filtering for these categories to gain a better understanding of how these categories are influencing Attrition.

For Linear Regressions, Monthly Income is significant at the 5% level as the constant and Total Working Years is signficant at the 1% level. You can filter the data to prioritize Total Working Years as a measure for what is influencing Attrition rates. The other variables that were found to be significant at the 1% level could also be filtered to better understand how those variables are influencing Attrition.

Attrition is a central topic in the modern workplace. The attrition rates are high and companies lose significant amounts of revenue from turnonver. The findings from this study would be very helpful for companies that would like to understand what is influencing these high levels of Attrition at their organizations. Narrowing down the findings that are statstically signficant or have high metric values for the models generated is a good place to start. The ultimate goal would be to retain more employees and increase revenue.