Health Catalyst data science take-home

Overview: In this exercise you will build simple machine learning models from the file called AWemployees.csv. Please complete the assignment using Jupyter notebook, RMarkdown, or similar. You will have two hours to complete the assignment, measured from the time you download it. Do not worry if you are unable to complete every analysis you would like to complete, just do as much as you can. When complete, please email your work (in an HTML or PDF file) to levi.thatcher@healthcatalyst.com and trevor.smith@healthcatalyst.com.

1. Using either R or Python on this VM, do the following and carefully document the steps in either a RMarkdown or Jupyter notebook.
    a. Connect to SQL Server and pull the following table into a dataframe: AdventureWorks2012.HumanResources.Employee
    b. Using the other columns as features, predict the SalariedFlag column using two common supervised learning methods.

2. Be sure to demonstrate the following:
    - Discuss which columns (if any) should be transformed or removed
    - Do a grid search over two common hyperparameters for each method
    - Discuss whether tuning hyperparameters helped at all vs the defaults
    - Avoid over-fitting the model
    - Provide appropriate performance estimates for the methods/models chosen
    - Discuss the differences in performance between the models
    - Train the better model on the entire data set using the best hyperparameters found.

Please work all comments into the same notebook (ie, with the code).

Thanks!