ITCS159 Software Lab for Basic Scientific Problem Solving
Lab 11: Introduction to data science using case studies in Kaggle

October 25, 2021

Objective

According to WiKipedia, `data science` is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured (e.g. texts posted on social networks). In this decade, data scientist becomes the sexiest job of the 21st Century[1]. A data scientist is a professional responsible for collecting, analyzing and interpreting large amounts of data to identify ways to help a business improve operations and gain a competitive edge over rivals. Basic responsibilities include gathering and analyzing data, and using various types of analytics and reporting tools to detect patterns, trends and relationships in data. Data scientists typically work in teams to mine big data for information that can be used to predict customer behavior and identify business risks and opportunities.

This lab aims to introduce the tasks of a data scientist via `Kaggle`. Kaggle is the world's largest community of data scientists and machine learners, owned by Google, Inc. Kaggle got its start by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and short form AI education. You can explore Kaggle via `https://www.kaggle.com/`.

Today, we will use a dataset provided by Kaggle for our study. We will go through the practices of a data scientist to analyze and extract insightful information from the dataset.

Getting to know your dataset

The dataset that you are going to work with relates to Human Resources. Read the following story to understand your analysis goals:

[1] `https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century`

Human Resources Analytics

*"Yeah, they all said that to me...", *Bob replied as we were at Starbucks sipping on our dark roast coffee. Bob is a friend of mine and was the owner of a multi-million dollar company, that's right, "m-i-l-l-i-o-n". He used to tell me stories about how his company's productivity and growth has sky rocketed from the previous years and everything has been going great. But recently, he's been noticing some decline within his company. In a five month period, he lost onefifth of his employees. At least a dozen of them throughout each department made phone calls and even left sticky notes on their tables informing him about their leave. Nobody knew what was happening. In that year, he was contemplating about filing for bankruptcy. Fast-forward seven months later, he's having a conversation with his cofounder of the company. The conversation ends with, "I quit..."*

That is the last thing anybody wants to hear from their employees. When a company experiences a high rate of employee turnover (They are quit!!!), then something is going wrong. This can lead the company to huge monetary losses by these innovative and valuable employees.

Companies that maintain a healthy organization and culture are always a good sign of future prosperity. Recognizing and understanding what factors that were associated with employee turnover will allow companies and individuals to limit this from happening and may even increase employee productivity and growth.

The objective of this analysis is that the company wants to understand what factors contributed most to employee turnover and to create a model that can predict if a certain employee will leave the company or not. The goal is to create or improve different retention strategies on targeted employees. Overall, the implementation of this model will allow management to create better decision-making actions.

Now, open the Jupyter Notebook Application. Then, go to Desktop and create folder here named `Lab11`.

Click `New` > `Python3` with name `lab11`. Now you can start the following exercise.

Obtaining the data

Let's start! you have to download the dataset from Mycourse called `HR_comma_sep.csv` and put it in the same folder as your notebook file (lab11).

and import the following library:

```
# Import the neccessary modules for data manipulation and
    visual representation
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib as matplot
import seaborn as sns
%matplotlib inline
```

Note that if you found errors saying that **No module**, you can use conda installation command to install missing packages. For example, you can use `conda install matplotlib` to install `matplotlib`.

Next, Read the csv file that you have downloaded from Mycourse and store our dataset into a dataframe called `df`.

```
df = pd.read_csv('HR_comma_sep.csv', index_col=None)
```

Scrubbing the Data

Typically, cleaning the data requires a lot of work and can be a very tedious procedure. This dataset from Kaggle is super clean and contains no missing values. But still, you have to examine the dataset to make sure that everything else is readable and that the observation values match the feature names appropriately.

Use the following command to check whether there are any missing values in the dataset

```
df.isnull().any()
```

What is the output? Try to interpret the output.

Use the following command to get a quick overview of what you are dealing with in the dataset.

```
df.head()
```

What is the output? Try to interpret the output.

As you can see, the column names are too long. Use the following code to rename certain columns for better readability.

```
df = df.rename(columns={'satisfaction_level': 'satisfaction',
        'last_evaluation': 'evaluation',
        'number_project': 'projectCount',
        'average_montly_hours': 'averageMonthlyHours',
        'time_spend_company': 'yearsAtCompany',
        'Work_accident': 'workAccident',
        'promotion_last_5years': 'promotion',
        'sales' : 'department',
        'left' : 'turnover'
        })
df.head()
```

Now, you need to move the **turnover** column to the front of the table.

```
front = df['turnover']
df.drop(labels=['turnover'], axis=1,inplace = True)
df.insert(0, 'turnover', front)
df.head()
```

What is the output? Try to interpret the output.

Milestone 1: You have now completed the Milestone 1. Keep going!

Exploring the Data

You will explore the dataset in terms of statistical characteristics. You can use the following commands to answer the questions:

```
print(df.shape)

print(df.dtypes)

df.describe()
```

How many rows and column contain in the dataset?

What is the data type of the features (columns)?

```
turnover
satisfaction
evaluation
projectCount
averageMonthlyHours
yearsAtCompany
workAccident
promotion
department
salary
```

What is the mean of satisfaction values?

What is the maximum value of yearsAtCompany?

What is the standard deviation value of averageMonthlyHours?

Now, it is time to explore turnover rate. Use the following command and try to interpret the output.

```
turnover_rate = df.turnover.value_counts() / len(df)
turnover_rate
```

What is the meaning of the output?

Now, let's explore the overview of turnover vs non-turnover in terms of the mean value of each feature.

```
turnover_Summary = df.groupby('turnover')
turnover_Summary.mean()
```

What is the meaning of the output?

Milestone 2: You have now completed the Milestone 2. Keep going!

Correlation Matrix and Heatmap

We will now explore the statistical characteristics using diagrams rather than just numbers to get more insight.

Use the following code to draw a correlation matrix and heatmap:

```
corr = df.corr()
sns.heatmap(corr,xticklabels=corr.columns.values,yticklabels=
    corr.columns.values,cmap="BuGn")
corr
```

What is the meaning of the output?
What is the different between positive and negative numbers?
What features affect our target variable the most (turnover)?
What features have strong correlations with each other?

Salary V.S. Turnover

Use the following code to draw a diagram showing the analysis between salary and turnover rate to answer the following questions.

```
f, ax = plt.subplots(figsize=(15, 4))
sns.countplot(y="salary", hue='turnover', data=df).set_title('
    Employee Salary Turnover Distribution')
```

Which salary level is the majority of employees who left the company?
Are there any high salary employee left the company?

Department V.S. Turnover

Let's see more information about the departments and answer the questions.

```
f, ax = plt.subplots(figsize=(15, 5))
sns.countplot(y="department", hue='turnover', data=df).
    set_title('Employee Department Turnover Distribution')
```

What are the top three highest turnover rate departments?

Turnover V.S. ProjectCount

Use the following code to draw a diagram showing the analysis between ProjectCount and turnover rate to answer the following questions.

```
ax = sns.barplot(x="projectCount", y="projectCount", hue="
    turnover", data=df, estimator=lambda x: len(x) / len(df) *
    100)
ax.set(ylabel="Percent")
```

Can you try to summarize some interesting points from the figure e.g. (what is the majority of the employees who did not leave the company)?

Turnover V.S. Evaluation

Let explore the correlation between the employee evaluation and turnover. Use the following to draw Density plot.

```
fig = plt.figure(figsize=(15,4),)
ax=sns.kdeplot(df.loc[(df['turnover'] == 0),'evaluation'] ,
    color='b',shade=True,label='no turnover')
ax=sns.kdeplot(df.loc[(df['turnover'] == 1),'evaluation'] ,
    color='r',shade=True, label='turnover')
ax.set(xlabel='Employee Evaluation', ylabel='Frequency')
```

```
plt.title('Employee Evaluation Distribution - Turnover V.S. No
    Turnover')
```

Which evaluation level of employees tend to leave the company?

Milestone 3: You have now completed the Milestone 3. Raise YOUR HAND NOW!