ANKARA UNIVERSITY

ENGINEERING FACULTY

DEPARTMENT OF COMPUTER ENGINEERING

INTERNSHIP REPORT

Data Science with Machine Learning and Deep Learning

Fatıma Rabia Yapıcıoğlu

16290130

10.08.2019

ABSTRACT

The objective of this summer internship report is to generally present an overview of data science with machine learning techniques currently in use or taken into consideration by data scientists worldwide. Aim of this internship is first to learn about data science, big data visualization, machine learning and deep learning, implement the following tools: anaconda, jupyter notebook, spider and the libraries following: numpy,pandas,matplotlib,seaborn,plotly,scikit-learn by describing the platforms used by data scientists like kaggle and finally integrating the machine learning into e-government platform. Since these terms mentioned above generally used in others place wrongly and create confusion in people's mind, each of them explained section by section. During this internship, more than one projects created in the fields of big data visualization and machine learning which aim to improve e-government platform by using new technologies. As a result, I've combined my old software engineering skills with this new data science experience and I am able to write new kernels and join kaggle's machine learning competitions and forum discussions which are very useful for my career as a data scientist. At the end of my internship, all of these subjects and projects have been presented in my internship presentation which has graded (scored) by engineers of Türksat Bilişim.

INSTITUTION INFORMATION

Name : Türksat Uydu Haberleşme Kablo TV ve İşletme A.Ş

Deparment : E-devlet ve Bilgi Toplumu Direktörlüğü

Address : Türksat Bilişim Bahçelievler Mahallesi 319 Cadde No:35/4(Ek Kuluçka Binası) 06830 Gölbaşı/ANKARA

Telephone : +90 312 925 59 00

E-mail : info@turksat.com.tr

Web Page (if exists) : https://bilisim.turksat.com.tr

Türksat Uydu Haberleşme Kablo TV ve İşletme A.Ş. (Türksat Company) is one of the world's leading companies providing all sorts of satellite communications through the satellites of Türksat as well as the other satellites. Providing services for voice, data, internet, TV, and radio broadcasting through the satellites across a wide area extending from Europe to Asia. Türksat Bilişim brand of Türksat, which has been working in satellite and cable services for many years, set out with the goal of becoming the Public Information Solution Center in 2011. Türksat Bilişim, which produces sustainable and high quality services with the assurance of Türksat, has successfully completed the IT projects needed by the public sector and closed an important gap in there.

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background and Subjects Declarations

During my internship, I got a chance to work in the department (E-devlet ve Bilgi Toplumu Direktörlüğü) of Türksat to know about how a software company uses big data in applications used by various public institutions,so the department which I was working on naturally dealing with massive volume of data that concerns data science.

- **Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract value from data.

Data scientists combine a range of skills—including statistics, computer science, and business knowledge,machine learning, deep learning —to analyze data collected from the web, smartphones, customers, sensors, and other source. These data sets are so voluminous that traditional data processing software just can't manage them,so big data is becoming one of the most important technology trends that has the potential for dramatically changing the way organizatins use information to enhance the customer experience and transform their models. For example,

- **Big Data visualization** is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

This mass of data is useless unless we analyse it and find the patterns hidden within.

- **Machine Learning** techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover.We'll classify it supervised and unsupervised learning and we'll implement the reccommender system of e-government platform.

1.2. Outline of the Report

Section II : DATA SCIENCE

- **I.** Definitions and Declarations about Data Science
- **II.** Tools and Platforms for Data Scientists
- **III.** Data Science Modeling Steps Used During Internship
- **IV.** How Data Science Is Transforming Business ?
- **V.** How Data Science Is Conducted ?
- **VI.** Data Science and the Growth of Data
- **VII.** Definition of Big Data
- **VIII.** Characteristics of Big Data
- **IX.** How Big Data Works ?
- **X.** Big Data Visualization
- **XI.** Hands-on Big Data Visualization
- **XII.** Projects

Section III: MACHINE LEARNING

- **I.** Definitions and Declarations about Machine Learning
- **II.** Regression Algorithms
- **III.** Classification Algorithms
- **IV.** Recommendar Systems
- **V.** Hands-on Machine Learning
- **VI.** Projects

Section IV:OUT OF TOPIC INTERN ACTIVITIES OF TÜRKSAT

- **I.** First Day of Intern:Get to know about E-devlet ve Bilgi Toplumu Direktörlüğü
- **II.** Panel:Get to know about Türksat
- **III.** E-devlet Anahtar Application Explatory Metarial
- **IV.** CİMER message classification
- **V.** Reinforcement Learning Presentation Of Ümit Köse(BT Engineer)

2. DATA SCIENCE

2.1. Definitions and Declarations about Data Science

Data science is the future of Artificial Intelligence.Therefore, it is very important to understand what is Data Science and how can it add value to your business. Traditionally,the data that we had was mostly structured and small in size,which could be analyzed by using the simple tools like SQL,PostgreSQL,Oracle, etc. Unlike data in the traditional systems which was mostly structured, today most of the data is unstructured or semi-structured.This data is generated from different sources like financial logs,text files,multimedia forms,sensors and instruments.Simple tools mentioned above are not capable of processing this huge volume and variety of data.This is why we need more complex and advanced analytical tools and algorithms for processing,analyzing and drawing meaningful insights out of it.

Data science reveals trends and produces insights that businesses can use to make better decisions and create more innovative products and services. Data is the bedrock of innovation, but its value comes from the information data scientists can glean from it and then act upon.

2.2. Tools and Platforms for Data Scientists

2.2.1. Data Collection Tools

Collecting quality data that can be transformed into rich analysis is the starting point every data strategy.The right data collection tools can reduce errors and duplicates,ensure greater accuracy,and preserve the integrity of data coming from all sources.

- GoSpotCheck,IBM Datacap,Mozenda,Octoparse, etc.

2.2.2. Data Analysis Tools

Finding meaning in and extracting value from your data is the core of all data analysis tools that enable you to easily understand and derive real meaning from your data help you male right business decisions that impact revenue and competitiveness.

- Alteryx,Domino Data Lab,KNIME Analytics Platform, etc.

2.2.3. Data Warehousing Tools

Data warehouses function as repositories for data tha's been combined and integrated from multiple,disparate sources and then standardized for ease of use.

- Amazon RedShift,Google BigQuery,Microsoft Azure,MySQL, etc.

2.2.4. Data Visualization Tools

Visual analytics tools identify patterns and trends in your data and help end users Understand and digest complex concepts.

- Google Fusion Tables ,Microsoft Power BI,SAS, etc.

2.2.5. Data Scientists Home: Kaggle

Kaggle is an online community of data scientists and machine learners, owned by Google LLC. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment work with other data scientists and machine learning engineers and enter competitions to solve data science challenges.

2.3. How Data Science Is Transforming Business ?

Organizations are using data science teams to turn data into a competitive advantage by refining products and services. For example, companies analyze data collected from call centers to identify customers who are likely to churn, so marketing can take action to retain them. Logistics companies analyze traffic patterns, weather conditions, and other factors to improve delivery speeds and reduce costs. Healthcare companies analyze medical test data and reported symptoms to help doctors diagnose diseases earlier and treat them more effectively.

2.3.1. How Data Science is Conducted ?

The process of analyzing and acting upon data is iterative rather than linear, but this is how the work typically flows for a data modeling project:

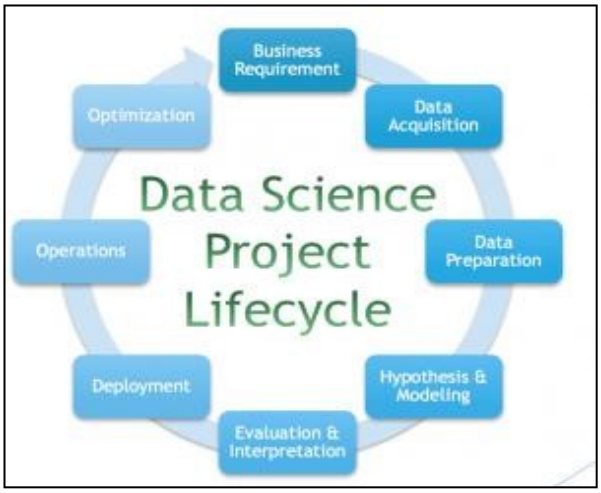


Figure 2.1. Data Science Project Lifecycle

2.3.2. Data Science and the Growth of Data

As modern technology has enabled the creation and storage of increasing amounts of information, the volume of data has soared. It's estimated that 90 percent of the data in the world was created in the last two years. For

example, Facebook users upload 10 million photos every hour. The number of connected devices in the world—the Internet of Things (IoT)—is projected to grow to more than 75 billion by 2025. The wealth of data being collected and stored by these technologies can bring transformative benefits to organizations societies around the world but only if we can interpret it. That's where data science comes in.

2.4. Big Data and Big Data Visualization

2.4.1. Big Data Definition and Characteristics

Big data is becoming one of the most important technology trends that has the potential for dramatically changing the way organizatins use information to enhance the customer experience and transform their models. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before.

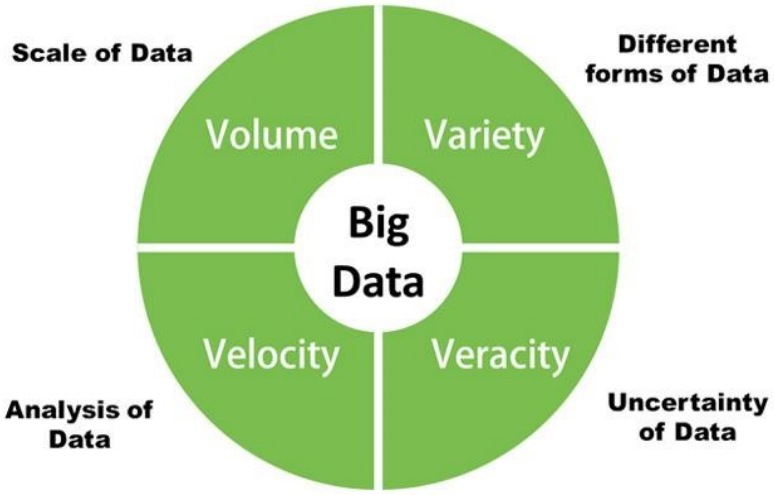


Figure 2.2. Characteristics of Big Data

2.4.2. Big Data Visualization

Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments.Data visualization can also:

- Identify areas that need attention or improvement.
- Clarify which factors influence customer behavior.
- Help you understand which products to place where.
- Predict sales volumes.

2.5. Hands-on Big Data Visualization

In this section, I'll show you how to investigate a dataset from the very start setp by step before projects part.

1.Import Required Libraries

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import matplotlib.pyplot as plt
from collections import Counter
%matplotlib inline
```

Figure 2.3. How to import required libraries for visualization ?

In this report I'll use python programming language and for visualization python has matplotlib,seaborn,plotly, etc. libraries.Here in the above code segment we imported the required libraries for visualization.Numpy is imported for linear algebra operation,Pandas for file operations,Seaborn and Matplotlib for visualization operations and others for structural requirements.

2.Reading the csv Files From Directory

```
kill = pd.read_csv('../input/PoliceKillingsUS.csv', encoding="windows-1252")
```

Figure 2.4. read_csv() method usage to read file from true directory.

We may have one or more datasets in separate files so they must be read by using pandas library one by one.Here hte method is read_csv used for file reading.And the kill variable is holding the dataframe after reading this file.

3.Understanding The Data

```
In [6]: kill.head()
```

Out[6]:

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_illness	threat_level
0	3	Tim Elliot	02/01/15	shot	gun	53.0	M	A	Shelton	WA	True	attack
1	4	Lewis Lee Lembke	02/01/15	shot	gun	47.0	M	W	Aloha	OR	False	attack
2	5	John Paul Quintero	03/01/15	shot and Tasered	unarmed	23.0	M	H	Wichita	KS	False	other
3	8	Matthew Hoffman	04/01/15	shot	toy weapon	32.0	M	W	San Francisco	CA	True	attack
4	9	Michael Rodriguez	04/01/15	shot	nail gun	39.0	M	H	Evans	CO	False	attack

Figure 2.5. Examining the first 5 rows of sample dataset with head().

Ater reading the file we must understand the data,understanding the column names and datasets story is import for exploratory data analysis (EDA) .By using head() method we can see the first five rows of the dataset, and by using tail() we can see the last five columns.

4.Data Preparation and Cleaning

```
# Race rates according in kill data
kill.race.dropna(inplace = True)
```

Figure 2.6. Examining dropping the null values with dropna() method.

Data preparation and cleaning part is also important to meaningful exploratory data analysis.For example, here with dropna(inplace=True) command we drop the null values permanently.

5.Visualization Tools

We have a lot of visualization tools in different libraries but just to demonstrate we'll use matplotlib for understanding types of charts.Some of the types we use in visualization part are like the following,

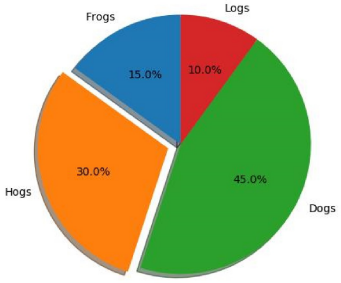


Figure 2.7. Pie Chart

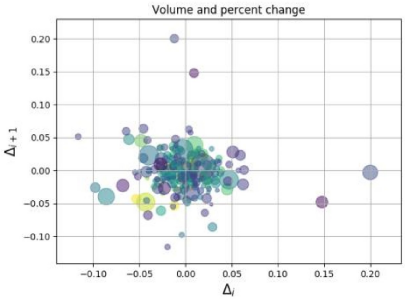


Figure 2.8. Scatter Plot

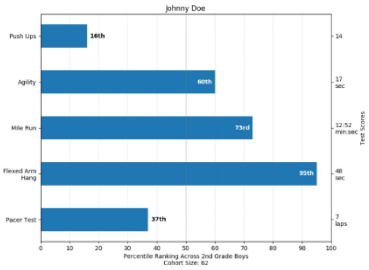


Figure 2.9. Barplot

2.6. Related Project During Internship

Name: What is the Turkish children's place in statistics ?

Subject: Exploratory Data Analysis(EDA),Data Visualization

Aim: Prepare and visualize the datasets of **TUIK(Turkish Statistical Institute)**

Summary of Project:

In this project raw data is imported from TUIK but then this dataset is cleaned and formatted according to the appropriate structural types.After prepration and cleaning the dataset we convert it into the csv file.We worked with more than one csv file that each of them is about Turkish children's.Aim of this project is to get a summary and visualization of the dataset which is also created by me.

Related Dataset Example We Have Used 1:

	Year	Total population	Total child population	Proportion Of child Population in total population
0	1935	16158018	7277722	45.0
1	1945	18790174	8667314	46.1
2	1950	20947188	9470412	45.2
3	1955	24064763	10902635	45.3
4	1960	27754820	12823514	46.2

Figure 2.10. Total Population-Total Child Population

By using this Total Population vs. Total Child Population dataset we can see the proportion of child population in total population year by year.And we can visualize this dataset with plotly library of python for example as following,

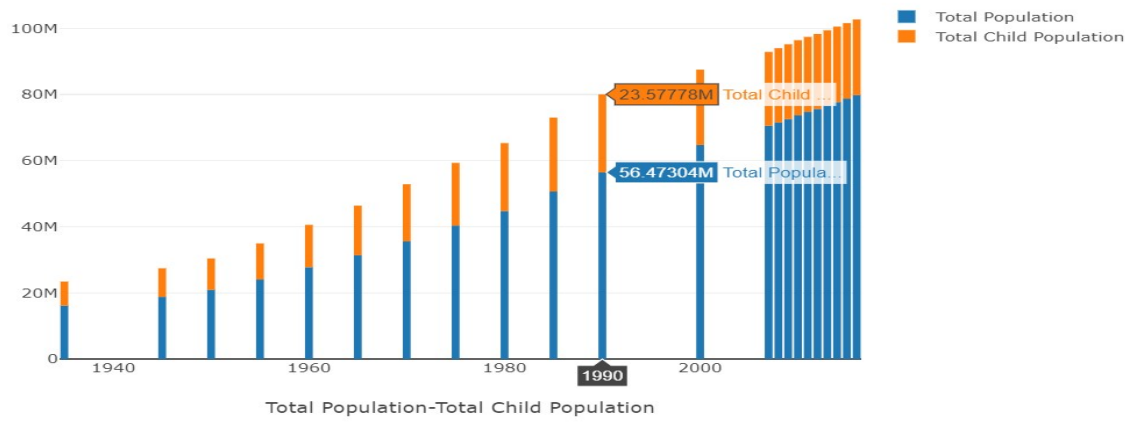
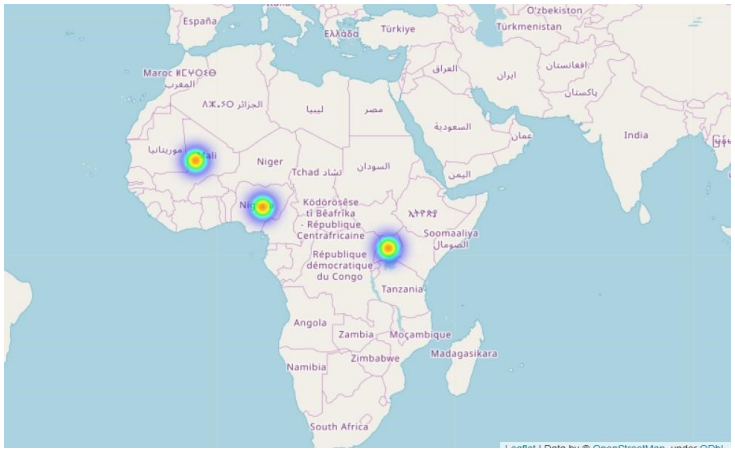


Figure 2.11. Relative Bar Charts

Also by using folium library we can visualize the top 3 countries according to the column of proportion of child population in total population on the world map as,



We can see the **top 3** countries which have the highest Proportion of Child population they are **Nigeria, Uganda,Mali** in order.

Figure 2.12. World Map with Folium Library

For example Let us see how we created world map by which code segment is as ,

```
import folium
import plotly.graph_objs as go
import folium.plugins as plugins

lat=16.4499982
lang= 14.5166646

world=folium.Map(location=[lat,lang],zoom_start=4)

dataUsa=np.array([[9.077751, 8.6774567,earth['childpopulation'][0]],[17.5739347, -3.9861092,ear
th['childpopulation'][1]],[1.3707295, 32.3032414,earth['childpopulation'][2]]])

plugins.HeatMap(dataUsa,name='Highest Number Of Childs',radius=15).add_to(world)
```

Figure 2.13. World Map with Folium Library

3. Machine Learning

3.1. What Is Machine Learning?

In the past 30 years there has been an explosion of data.This **mass of data is useless unless we analyse it** and find the patterns hidden within.Machine Learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover.The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making.

3.2. Two Approaches To Machine Learning

3.2.1 Supervised Machine Learning

Supervised machine learning algorithms are the most commonly used. With this model, a data scientist acts as a guide and teaches the algorithm what conclusions it should make. Just as a child learns to identify fruits by memorizing them in a picture book, in supervised learning, the algorithm is trained by a dataset that is already labeled and has a predefined output.Examples of supervised machine learning include algorithms such as linear and logistic regression, multiclass classification, and support vector machines.

3.2.2 Unsupervised Machine Learning

Unsupervised machine learning uses a more independent approach, in which a computer learns to identify complex processes and patterns without a human providing close, constant guidance. Unsupervised machine learning involves training based on data that does not have labels or a specific, defined output.To continue the childhood teaching analogy, unsupervised machine learning is akin to a child learning to identify fruit by observing colors and patterns, rather than memorizing the names with a teacher’s help.

3.3 Recommender Systems,User-Based and Item-Based Collaborative Filtering

3.3.1 User Based Collaborative Filtering

Collaborative filtering is making recommend according to combination of your experience and experiences of other people.

- First we need to make **user** vs **item** matrix.
- Each **row is users** and each **columns are items** like movie, product or websites
- Secondly, computes **similarity scores between users**.
- Each row is users and each row is vector.
- Compute similarity of these rows (users).
- Thirdly, **find users who are similar to you based on past behaviours**
- Finally, it **suggests that you are not experienced before.**

3.3.2. Item Based Collaborative Filtering

Instead of focusing on users, we could focus on wchich services from all the options are more similar to what we know he enjoys. We could divide IB-CF in two sub tasks: Calculate similarity among the items: Correlation-Based Similarity,Calculation of Prediction .

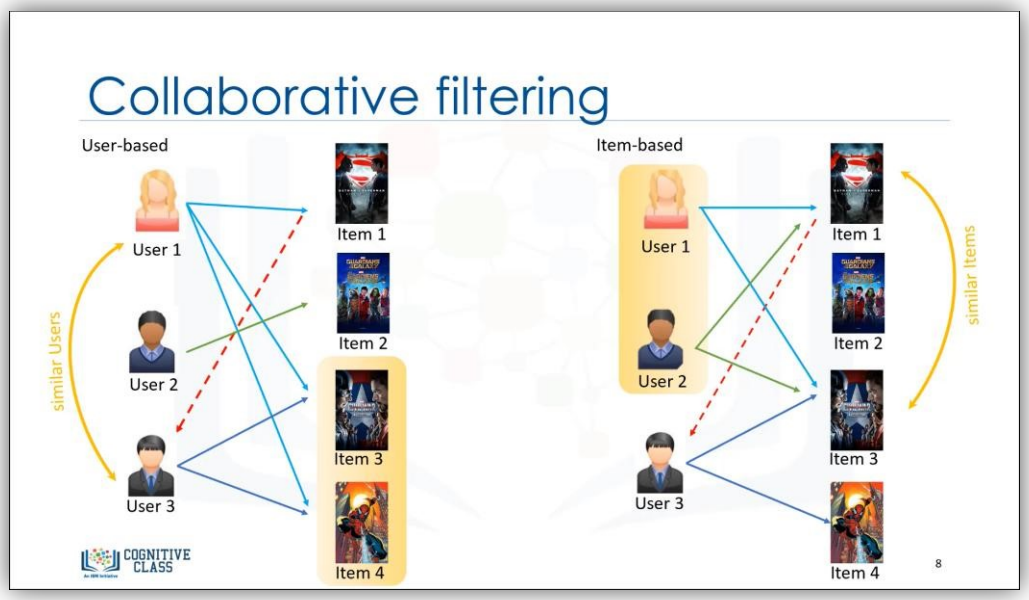


Figure 3.1. Collaborative Filtering Classification

3.4. Related Project During Internship

Name: E-Government,türkiye.gov.tr Supervised Recommender System

Subject: Machine Learning

Aim: Emerging Supervised Machine Learning's Recommender Algorithms to E-Government Platform

Summary of Project:

There are two types of machine learning algorithms: user based and item based collaborative filtering.In this projects we used item based collaborative filtering.In E-Government platform we have a lot of services that publicly provided.And this system makes some suggestions via this underlying machine learning algorithm.First we've prepared the sample dataset,

Service	Case File Inquiry	Criminal Record Inquiry	Criminal Record Verifying	Debt Inquiry	Individual Subscription Application	Interview Application For Judges and Prosecutors	Job Application	Receipt Information Inquiry	Subscriber Info Inquiry	natural gas subscription inquiry
UserId										
105555	NaN	NaN	NaN	NaN	4.6	NaN	NaN	NaN	NaN	NaN
115765	NaN	3.5	2.5	NaN	NaN	NaN	3.0	NaN	NaN	NaN
119965	3.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 3.2. Sample Dataset Of Users And Services

and then created a model which is an item based collaborative fitler and by using correlation based similarity.First,we are extracting the related service,

```
#Now let's grab the user ratings for those two movies:
Receipt_Information_Inquiry=moviemat['Receipt Information Inquiry']
```

Figure 3.3. Extracting Service From Dataset

After some other calculations,finally in the following code segment we calculate the correlation with corrwith() method,

```
#We can then use corrwith() method to get correlations between two pandas series:
similar_to_Receipt_Information_Inquiry=moviemat.corrwith(Receipt_Information_Inquiry)
```

Figure 3.4. corrwith() Method Usage

At the end we made a sort operation and we see the suggested service.

Service	Correlation
Individual Subscription Application	1.000000
Debt Inquiry	0.617415

Figure 3.5.

Suggested Service:Debt Inquiry

Project Plan Designation:

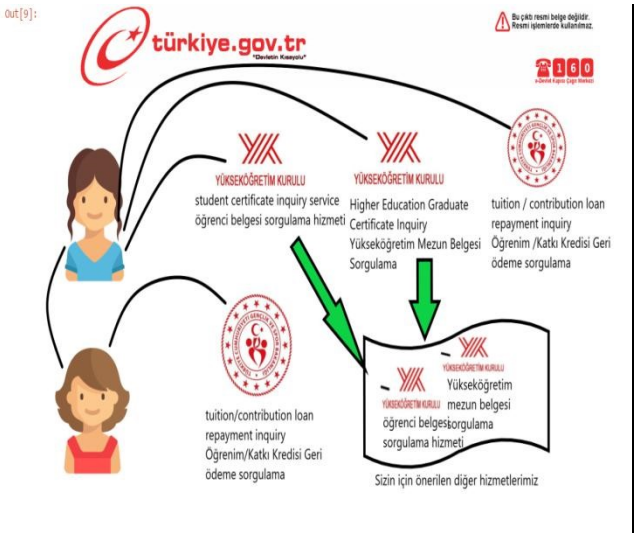


Figure 3.6. User Based C.F.

Figure 3.7. Item Based C.F.

4.OUT OF TOPIC ACTIVITIES DURING INTERNSHIP

4.1. First Day Activity: Introduction Of E-Government and Information Society Directorate

Figure 4.1. Intro Presentation

E-Government and Information Society provides E-Government services both in Web and Mobile Application.This platform focuses on citizens needs and citizens can attain the public institution services without going the institution itself with E-Government platform.E-government platform implement these services by applying

the institution when citizens requested on internet and doesn't keep the data itself. After that presentation interns chose the department and subject that they wanted to work with.

4.2. Türksat Promotion Panel

Figure 4.2. Panel

Türksat Human Sources scheduled a promotion presentation which aims to promote Türksat Company and departments of it ,business fields of Türksat for Interns.Also after presentation business and career conversation arranged by assistant general managers of Türksat.

Figure 4.3. Lagari Hasan Çelebi Satellite Museum.

Panel is concluded with Lagari Hasan Çelebi Satellite Museum.

All the satellites entering into Türksat inventory and models of launch rockets that take them into space are exhibited in the Lagari Hasan Çelebi Satellite Museum.

CONCLUSION

In this internship work, I've studied Data Science and Machine Learning techniques and especially Exploratory Data Analysis(EDA), Data Visualization, Supervised and Unsupervised learning algorithms, Recommender Systems and then implemented each subfield of my study with a new project but I gave place to only a few of these projects in this report. Aim of the studying Data Science is to learn what is the underlying information of Machine Learning. An important part of the studying Data Science is being a member of Kaggle which is platform known as Data Scientist's Home. Most of the Data Scientist is a member of this platform and improve their programming skills in the field of Machine Learning by using this platform.
In Data Visualization part, I've learned how to make data more understandable, by using different python programming language libraries. I also write some kernels which are about exploratory data analysis with different datasets and publish on Kaggle. And I've also learned how to create our own datasets and make their exploratory data analysis. When we are making exploratory data analysis we've used different platforms like anaconda's spider,jupyter notebook and Kaggle.In the second part of my studies, I focused on Machine Learning, I've taken courses from IBM, and Udemy. After learning machine learning techniques in general I also join the competitions of Kaggle which is in the field of Machine Learning to gain experience more.And finally I've learned how recommender systems work and implement it with e-government platform's sample dataset.

BIBLIOGRAPHY

Alley, G. , 2018. Top Data Science Tools, Alooma Blog,Data Analysis, https://www.alooma.com/blog/top-data-science-tools, Acces Date:11.08.2019

Edwards, G. 2018. Machine Learning, An Introduction, https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0, Acces Date:08.08.2019

Aghabozorgi, S. Cognitive Class IBM , Data Science and Machine Learning 101 Tut., https://www.youtube.com/watch?v=XzL8i_UqZyc&list=PLXeOa5hMEYxN1Kzrqlac6YQ8ULPJoexl

Palamada, H. , 2018. Data Visualization Handbook,Kaggle Kernels,https://www.kaggle.com/hiteshp/data-visualization-handbook/comments, Access Date: 12.08.2019

Banik, R. 2018. Recommender Systems in Python:Beginner Tutorial,Datacamp Blogs., https://www.datacamp.com/community/tutorials/recommender-systems-python

Anonymous. 2019. https://matplotlib.org/tutorials/index.html,Access Date:14.08.2019

Anonymous. 2019. https://www.turksat.com.tr/en/corporate/about-us

Yapıcıoğlu, R. 2019,What is the Turkish Children's place in statistics ?,Kernels, https://www.kaggle.com/rabiayapicioglu/dedicated-kernel-to-world-s-children-from-turkey, Access Date:14.08.2019

Patil, P. 2018. What is exploratory data analysis ? https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15, Access Date:15.08.2019

APPENDICES

Appendix 1. Data Science Terms

Exploratory Data Analysis (EDA):

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

EDA explained using sample Data set:

```
In [2]: df = pd.read_csv('winequality-white.csv',sep=';')
        df.head()
```

Out[2]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Original data is separated by delimiter " ; " in given data set.

To take a closer look at the data took help of " .head()"function of pandas library which returns first five observations of the data set.Similarly ".tail()" returns last five observations of the data set.We found out the total number of rows and columns in the data set using ".shape".

```
In [3]: df.shape
Out[3]: (4898, 12)
```

Dataset comprises of 4898 observations and 12 characteristics.

Out of which one is dependent variable and rest 11 are independent variables

physico-chemical characteristics.

```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
fixed acidity           4898 non-null float64
volatile acidity        4898 non-null float64
citric acid             4898 non-null float64
residual sugar          4898 non-null float64
chlorides               4898 non-null float64
free sulfur dioxide     4898 non-null float64
total sulfur dioxide    4898 non-null float64
density                 4898 non-null float64
pH                      4898 non-null float64
sulphates               4898 non-null float64
alcohol                 4898 non-null float64
quality                 4898 non-null int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

Data has only float and integer values.No variable column has null/missing values.The describe() function in pandas is very handy in getting various summary statistics.This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.

```
In [6]: df.describe()
```

Out[6]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	10.514267
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	1.230621
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	14.200000

Here as you can notice mean value is less than median value of each column which is represented by 50%(50th percentile) in index column.There is notably a large difference between 75th %tile and max values of predictors “residual sugar”,”free sulfur dioxide”,”total sulfur dioxide”.Thus observations 1 and 2 suggests that there are extreme values-Outliers in our data set.Few key insights just by looking at dependent variable are as follows:

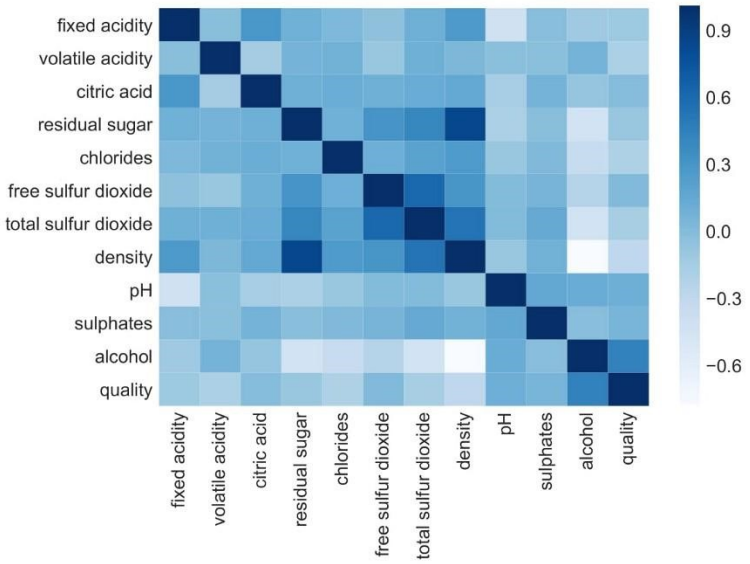
```
In [7]: df.quality.unique()
```

```
Out[7]: array([6, 5, 7, 8, 4, 3, 9], dtype=int64)
```

Target variable/Dependent variable is discrete and categorical in nature.“quality” score scale ranges from 1 to 10;where 1 being poor and 10 being the best.1,2 & 10 Quality ratings are not given by any observation. Only scores obtained are between 3 to 9.

```
In [8]: df.quality.value_counts()
Out[8]: 6    2198
        5    1457
        7     880
        8     175
        4     163
        3      20
        9       5
        Name: quality, dtype: int64
```

This tells us vote count of each quality score in descending order.“quality” has most values concentrated in the categories 5, 6 and 7.Only a few observations made for the categories 3 & 9.To use linear regression for modelling,its necessary to remove correlated variables to improve your model.One can find correlations using pandas “.corr()” function and can visualize the correlation matrix using a heatmap in seaborn.



CSV Files:

A Comma Separated Values (CSV) file is a plain text file that contains a list of data. These files are often used for exchanging data between different

applications. For example, databases and contact managers often support CSV files.

Numpy:

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

Pandas:

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. In this tutorial, we will learn the various features of Python Pandas and how to use them in practice.

Matplotlib:

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four

graphical user interface toolkits.

Plotly:

Plotly, also known by its URL, **Plot.ly**,[1] is a technical computing company headquartered in Montreal, Quebec, that develops data online dataanalytics and visualization tools. Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB, Perl, Julia, Arduino, and REST.

Seaborn:

Seaborn is a Python library created for enhanced data visualization. It's a very timely and relevant tool for data professionals working today precisely because effective data visualization – and communication in general – is a particularly essential skill. Being able to bridge the gap between data and insight is hugely valuable, and Seaborn is a tool that fits comfortably in the toolchain of anyone interested in doing just that.

Anaconda:

Anaconda is a python and R *distribution*. It aims to provide everything you need (python wise) for data science "out of the box".

It includes:

- The core python language
- 100+ python "packages" (libraries)
- Spyder (IDE/editor - like pycharm) and Jupyter
- conda, Anaconda's own package manager, used for updating Anaconda and packages