

광주 ICT 이노베이션스퀘어
인공지능 융복합교육

Semi-Project for Part 5

Training & testing traditional ML algorithms (Titanic survival analysis)

Daeyeon Jo
repositivator@gmail.com

본 교안은 광주 ICT 교육 과정을 위해 제작되었으며, 교육 외 배포/게시/공개를 금합니다.

Course Overview

* 아래 커리큘럼의 세부 사항은 변동될 수 있습니다.
* 진도 상황에 따라 1~2일 정도 차이가 발생할 수 있습니다.

2020년 10월

일

월

화

수

목

금

토

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

2020년 12월

일

월

화

수

목

금

토

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

파이썬 프로그래밍 기초

파이썬 정형 데이터 분석 (데이터 탐색 / 데이터 전처리 / 데이터 시각화)

파이썬을 활용한 데이터 수집 & 웹 스크레이핑 (+ 자동화 프로그램 개발)

파이썬 기반 텍스트 데이터 분석

Python 기반 기초 통계분석 (빈도분석 / 기술통계 / 교차검정 / 평균차이검정)

1차 세미 프로젝트 (데이터 수집 / 전처리 / 통계분석 / 시각화 + 팀별 발표)

SQL 기초 프로그래밍 (Data Modeling / SQL CRUD / Adv. Techniques)

머신러닝 핵심 이론 & 주요 알고리즘 이론

파이썬 기반 머신러닝 알고리즘 실습 (Scikit-learn)

+ 데이터 분석 관련 직무 & 학습 리소스 소개

2차 세미 프로젝트 (Feature engineering & applying ML algorithms)

딥러닝 핵심 이론 & 인공신경망 최적화 이론

파이썬 기반 딥러닝 알고리즘 실습 (Tensorflow & Keras)

+ 분야별 머신러닝 & 딥러닝 활용 사례 소개

파이널 프로젝트 (데이터 수집 / 탐색 & 전처리 / 시각화 + ML & DL model tuning)

최종 프로젝트 발표 & 수료식

수업 관련 공지사항

* 데이터 전처리 방법 / Model 선택 / Metric 선택 모두 자유입니다. (배운 내용의 복습에 Focus!)

* Part 1/2/4/5 에서 배운 지식들을 최대한 빠짐없이 활용하는데 초점을 맞춰주세요.

* “Titanic prediction” 등과 같이 관련 코드에 대한 직접적인 검색은 피해주세요.

* 발표 시 포함할 사항 : 데이터 전처리 방법 & 이유 / 모델 적용 프로세스 / 모델 적용 결과
발표 시 제출할 사항 : 상세 주석이 포함된 전체 코드 (.ipynb 제출, PPT 발표자료 필수 X)

* 화 ~ 수 : 팀별 분석 작업 -> 목요일 3~4교시 : 팀별 발표 및 질의응답 (15분 내외/팀)
: 목요일 2교시 종료 시까지 Jupyter notebook 제출 : 슬랙 DM or repositivator@gmail.com

* 1차 세미프로젝트 발표자는 발표 X & 도움이 필요할 경우 슬랙 채널에서 호출

수업 관련 공지사항

1팀 : 조혜빈, 안현진, 이승재, 한대영

2팀 : 강예은, 박광렬, 서경호, 정경훈

3팀 : 신지운, 강기훈, 문수지, 조우람

4팀 : 이하은, 고성헌, 김관주, 최갑주

* 아직 학습이 미진한 분들은 기존 Part 1 ~ Part 5의 내용들을 차례대로 실습해보고, 본인 팀과 다른 팀의 발표 자료를 통해 데이터에 모델을 적용하는 과정을 넓게 이해하는 것에 포커스를 맞춰주세요.

1. Blank notebook for this semi-project

ML for Titanic survival prediction (Blank)

Logout

FileEditViewInsertCellKernelWidgetsHelp

TrustedPython 3

Save

+

Undo

Copy

Paste

Up

Down

Run

Interrupt

Restart

Code

Terminal

In [1]:

```
1 import warnings
2 warnings.filterwarnings("ignore")
3
4 import pandas as pd
5 import numpy as np
6 import matplotlib.pyplot as plt
7
8 # from sklearn import ?
9 # from sklearn.metrics import ?
```

1. Preparing dataset (2번부터 실습)

In [12]:

```
1 data_df = pd.read_csv('titanic.csv')
2 data_df.head(3)
```

Data info

- **PassengerId** : Unique ID of passenger
- **Survived** : 0 = No, 1 = Yes
- **pclass** : Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **sibsp** : # of siblings & spouses aboard the Titanic
- **parch** : # of parents / children aboard the Titanic
- **ticket** : Ticket number
- **cabin** : Cabin number
- **embarked** : Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S

2. Possible pathways for data preprocessing

- + Check & adjust data for handling **Missing data & Outlier**
- + Select important columns (or just use all columns & improve your model later)
- + Change characters to numbers (Binary num, Class num, One-hot vector, etc.)
- + (If applicable & useful) **Select features** with Tree-based models
- + (If applicable & useful) **Modify the scale of features** with StandardScaler / MinMaxScaler
- + (If applicable & useful) **Reduce dimension** with PCA
- + (If applicable & useful) Try **other traditional ML Models** for enhancing the result (except DL/NN)

서울시 범죄현황 통계자료 분석 및 시각화

2. 서울시 범죄현황 통계자료 분석 및 시각화 Last Checkpoint: an hour ago (unsaved changes)

File Edit View Insert Cell Kernel Help Trusted

서울시 범죄현황 통계자료 분석 및 시각화

In [1]:
1 import numpy as np
2 import pandas as pd
3 import seaborn as sns
4
5 import matplotlib.pyplot as plt
6 from matplotlib import font_manager, rc

1. 데이터 입력 및 데이터 전처리

In [2]:
1 df = pd.read_excel('관서별 5대범죄 발생 및 검거.xlsx', encoding='utf-8')
2 df.head()

	관서명	소계(발생)	소계(검거)	살인(발생)	살인(검거)	강도(발생)	강도(검거)	강간(발생)	강간(검거)	절도(발생)	절도(검거)	폭력
0	계	126401	82688	163	156	276	257	5449	5069	55307	21842	652
1	충무서	2860	1716	2	2	3	2	185	65	1395	477	135
2	종로서	2472	1589	3	3	6	5	115	98	1070	413	127
3	남대문서	2094	1226	1	0	6	4	65	46	1153	382	869
4	서대문서	4029	2579	2	2	5	4	154	124	1812	738	205

Scikit-learn practices & Appendix

(Appendix 0) Missing data visualization (with missingno).zip

(Appendix 1) Tuning HyperParams with Hyperopt (+ LightGBM).zip

(Appendix 2) Auto-ScikitLearn (with Breast cancer data).zip

(Appendix 3) Auto-feature-engineering with FeatureTools.zip

(Appendix 4) PCA for BreastCancer & Cifar10 (딥러닝 학습 후 추가 학습).zip

(Appendix 5) Colab 기준 LightGBM 설치 방법 (2020.03).zip

1. (Cheat Sheet) Scikit_Learn.zip

2. Hands On MachineLearning with ScikitLearn and TensorFlow (1판 & 2판).zip

3. Model saving & loading (Scikit-learn).zip

4. Amazon SageMaker Hands-on.docx

5. AI Career Pathways with Roles, Tasks, Skills (Workera Report).pdf

파이썬을 활용한 기초 통계분석

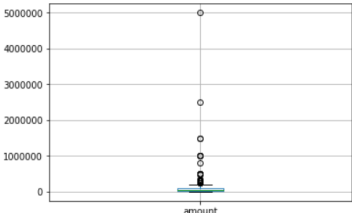
2. 빈도 분석 & 기술통계량 분석

File Edit View Insert Cell Kernel Widgets Help

3. Outlier 의 탐지 및 제거와 전후 분포 비교

In [218]:
1 df.boxplot(column='amount') # 위 아래의 작은 점(검정색 선이 상/하한선, 그 밖의 경우는

<matplotlib.axes._subplots.AxesSubplot at 0x1b1170839b0>

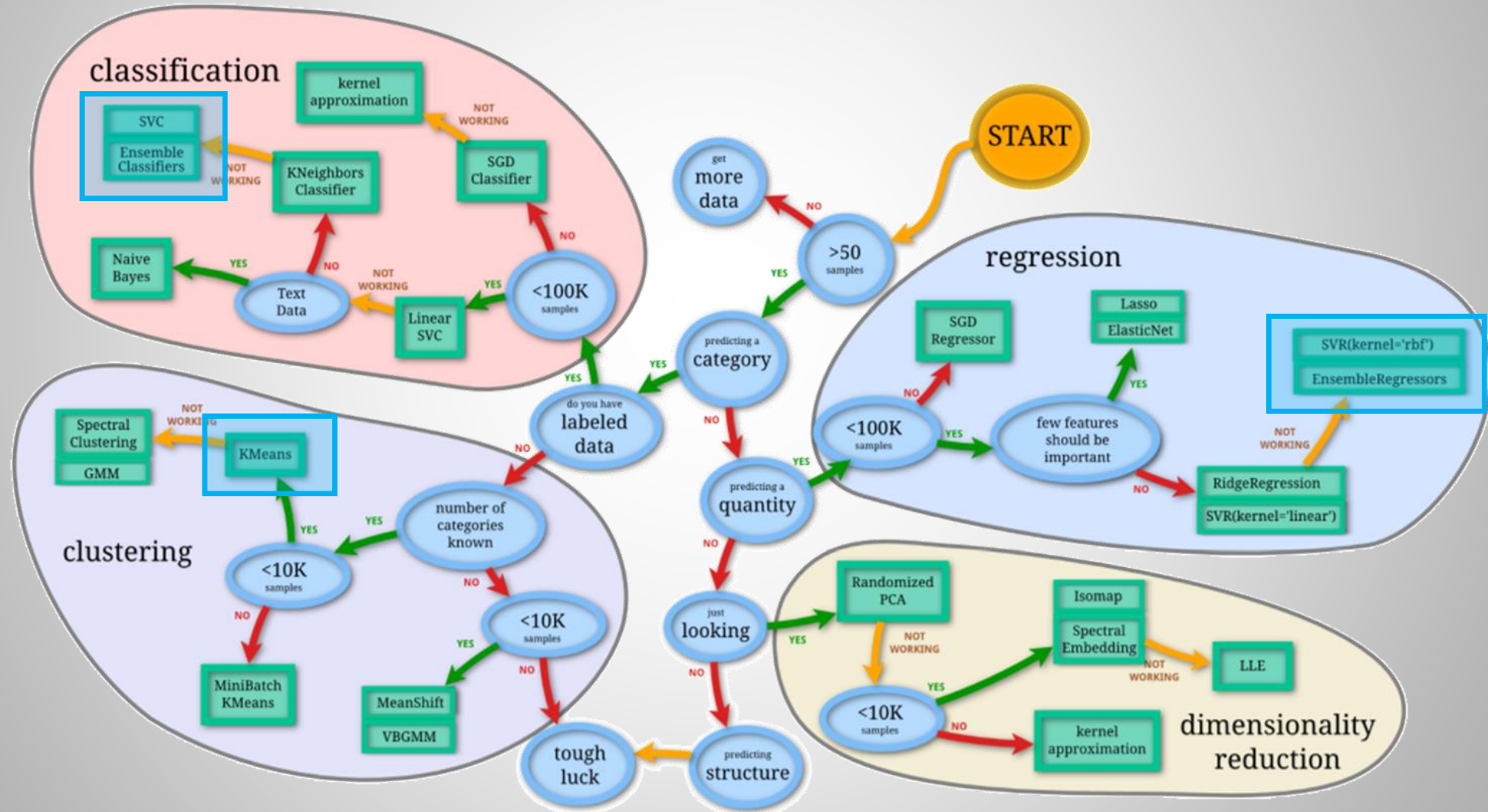


The boxplot displays the distribution of the 'amount' variable. The y-axis ranges from 0 to 5,000,000. The box represents the interquartile range (IQR) from approximately 0 to 1,000,000. The median is around 500,000. Whiskers extend to the minimum and maximum values within 1.5 times the IQR. Numerous outliers are plotted as individual points above the upper whisker, reaching up to 5,000,000.

3. Available traditional ML models (+ apply hyper-params optimization)

풀어내려는 문제의 종류와 데이터의 타입(형태, 수)에 따른 ML 알고리즘 선택 가이드

http://scikit-learn.org/stable/tutorial/machine_learning_map/ (각 알고리즘 별 예시 코드 有)



본 교안은 광주 ICT 교육 과정을 위해 제작되었으며, 교육 외 배포/게시/공개를 금합니다.

- + **Model stacking** 적용해보기
- + **AutoML** 적용해보기 (Google AutoML Tables, FeatureTools, Auto-sklearn 등)
- + **Bayesian Hyperparams Optimization** 적용해보기

광주 ICT 이노베이션스퀘어
인공지능 융복합교육

End of Document