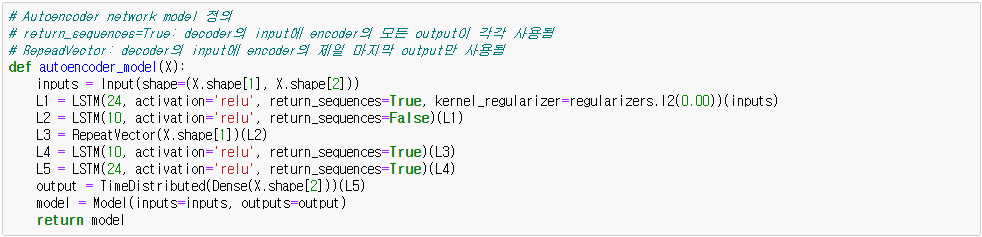
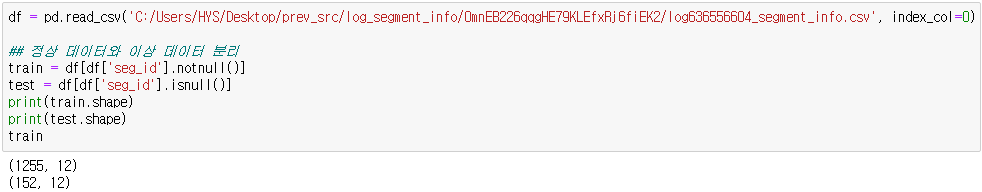
* **LSTM Autoencoder for Anomaly Detection**
  + **목표: 자동차는 가지 못하지만, 자전거는 갈 수 있는 길을 segment로 등록하자.**
  + **학습 전 문제점**
    - 무엇을 **이상치**로 규정해야 하는가?
      * **원하는 결과물**: segment.csv에 존재하지 않는 새로운 segment
      * **segment.csv**: 메르카토르 값(X1, Y1, X2, Y2), 위/경도(1, 2), tileId(X, Y)
        + 시계열 데이터가 아님
        + 시간의 흐름에 따라서 변화하는 데이터가 아니라, 단순한 정보의 모음
      * **log{logname}\_info.csv**: 메르카토르 값(X, Y), 위/경도, tileId(X, Y), 방향, 거리
        + segment와 관련하여 학습할 수 없음
      * **log{logname}\_segment\_info.csv**: 메르카토르 값(X, Y), 위/경도, tileId(X, Y), 방향, seg\_id, seg\_lat(1, 2), seg\_long(1, 2)
        + segment 매칭이 불가능한 경우: seg 관련 값이 모두 0

log 데이터와 segment 데이터가 함께 존재하기 때문에 학습 및 테스트용 데이터셋으로 사용하기에 가장 적합해 보임

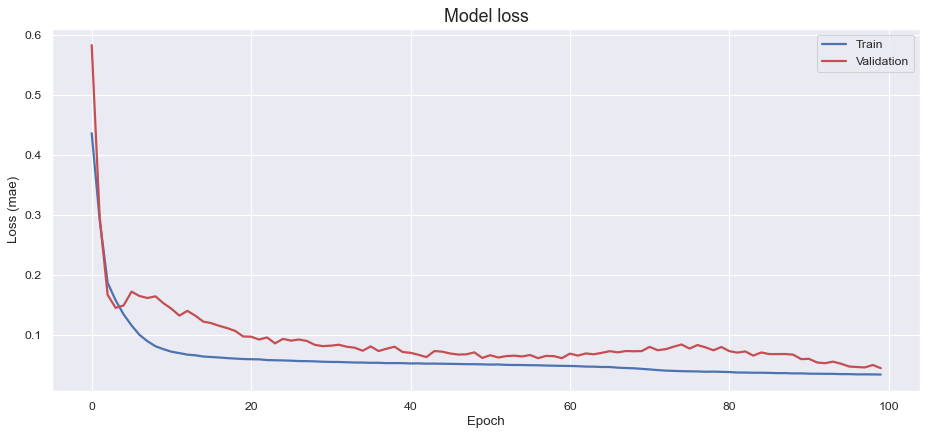
* + - 사용자의 GPS Log를 봤을 때, 잠깐 편의점을 들렸을 수도 있고, 길을 잘못 들었을 수도 있고 **경로를 이탈**하는 여러가지 경우의 수가 있는데, 이런 부분을 어떻게 제거해야 하는가?
      * 직접 추가한 segment 중 **사용 빈도수가 낮은 segment를 제외**하는 방식
    - 너무 가기 힘든, 지엽적인 길은 어떻게 제외할 것인가?
      * **최소한 인도, 차도는 깔려 있는 곳**을 segment로 사용하는 것이 좋을 것
* **1. LSTM Autoencoder for Anomaly Detection\_by Brent Larzalere**
  + Autoencoder 학습 과정에는 이상치 데이터가 없는 **정상 데이터만**을 사용
  + **학습 1**



* + - **데이터셋**: log636556604\_segment\_info.csv
      * **training set**: seg\_id != null인 row
        + matching된 segment가 존재하는 log들
      * **test set**: segid == null인 row



* + - **사용 변수**: 모든 변수(mercX, mercY, …, seg\_long2)
    - **결과:**

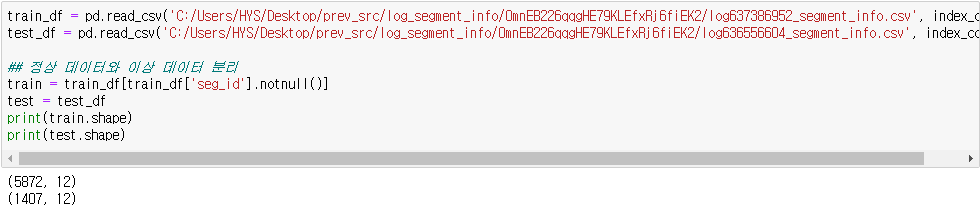




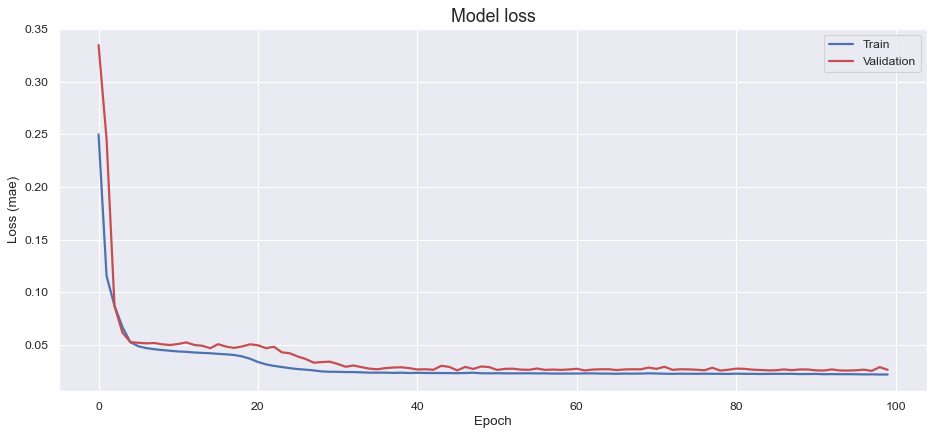
* + - * anomaly threshold = 0.09866794878754732

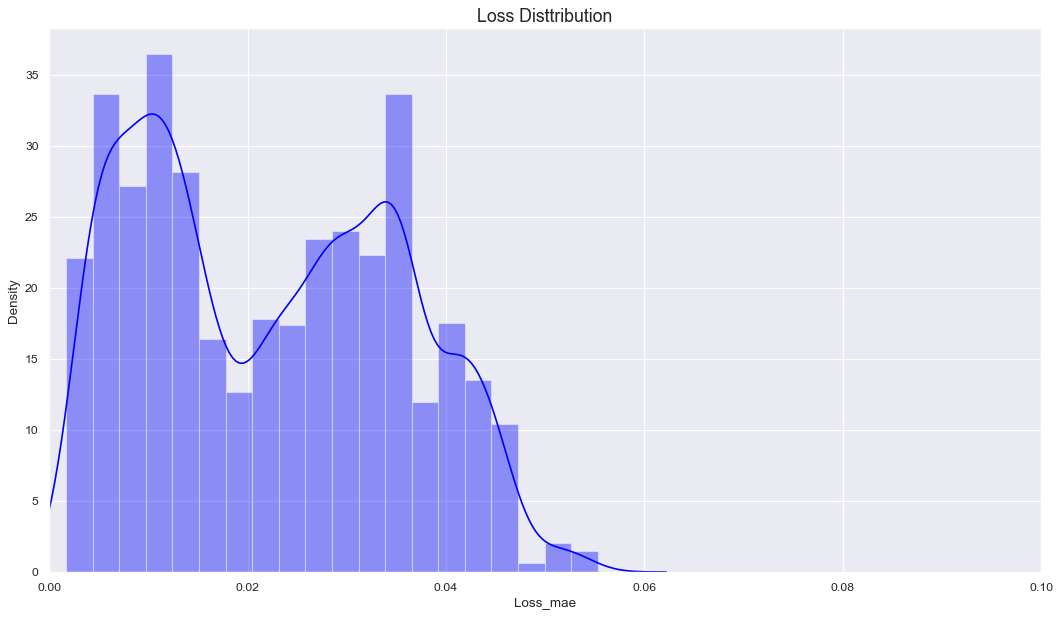


* + - * 이상치 없음
  + **학습 2**
    - **데이터셋**: log636556604\_segment\_info.csv, log637386952\_segment\_info.csv
      * **training set**: log637386952\_segment\_info.csv 중 segid != null
      * **test set**: log636556604\_segment\_info.csv

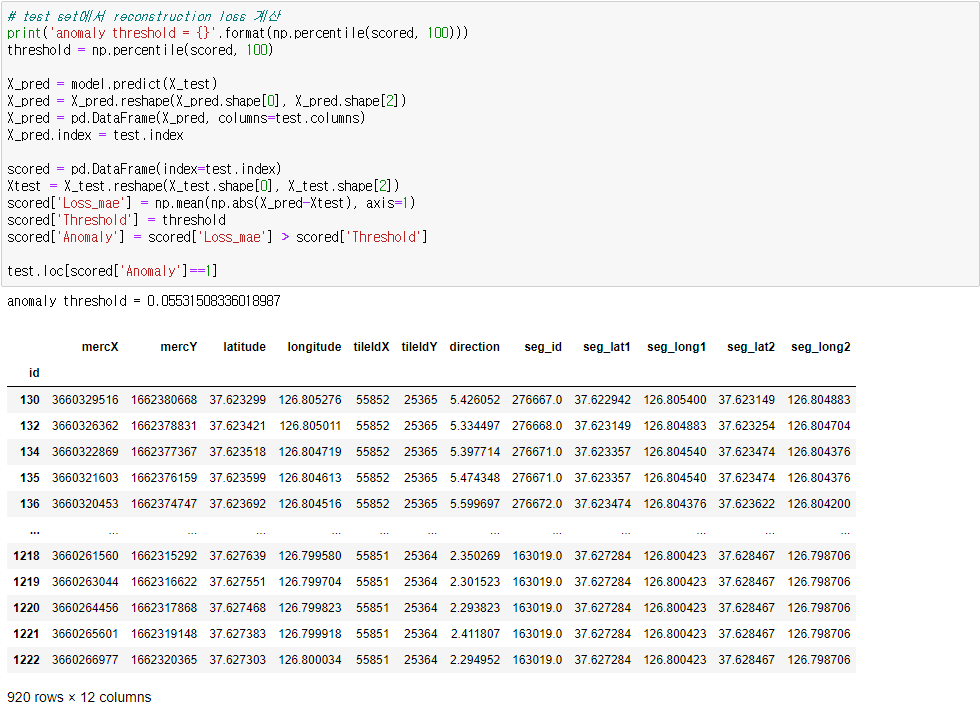


* + - **사용 변수**: 모든 변수(mercX, mercY, …, seg\_long2)
    - **결과:**

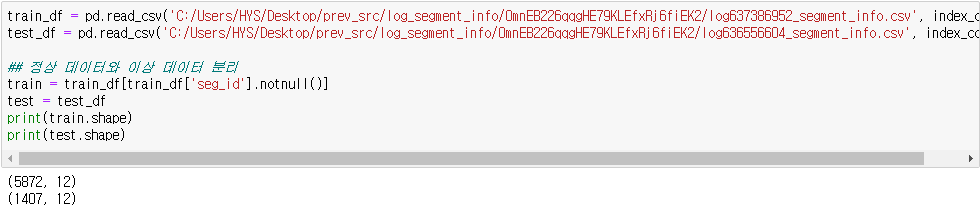




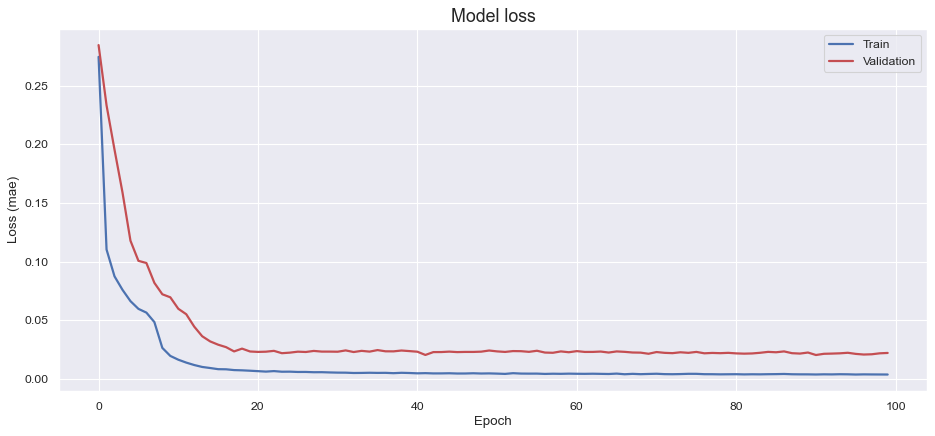
* + - * anomaly threshold = 0.05531508336018987

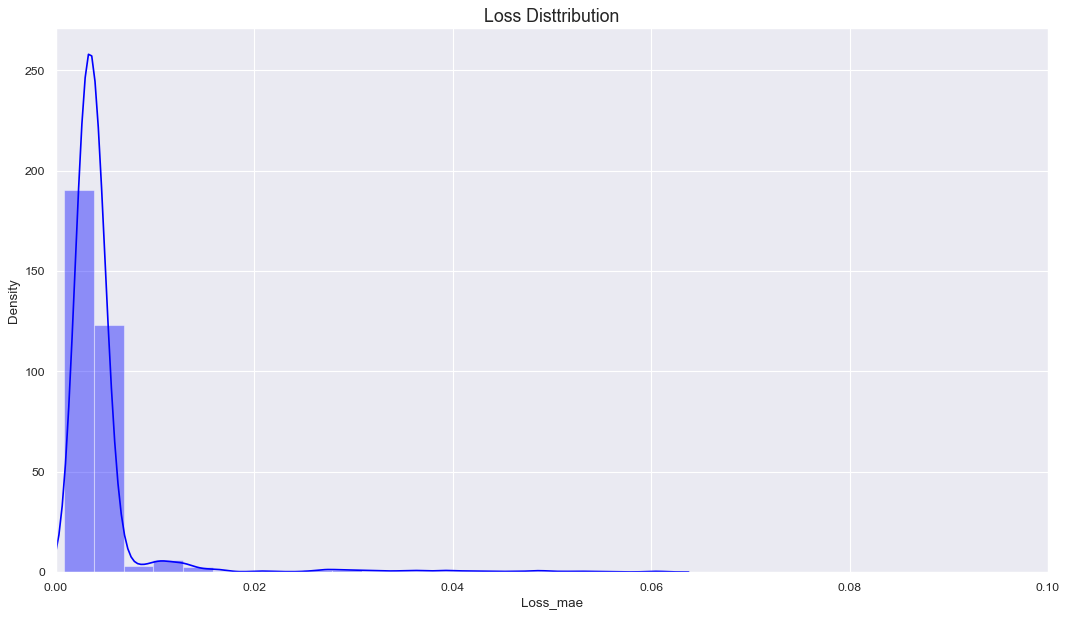


* + - * 920개의 이상치 발견
  + **학습 3**
    - **데이터셋**: log636556604\_segment\_info.csv, log637386952\_segment\_info.csv
      * **training set**: log637386952\_segment\_info.csv 중 segid != null
      * **test set**: log636556604\_segment\_info.csv

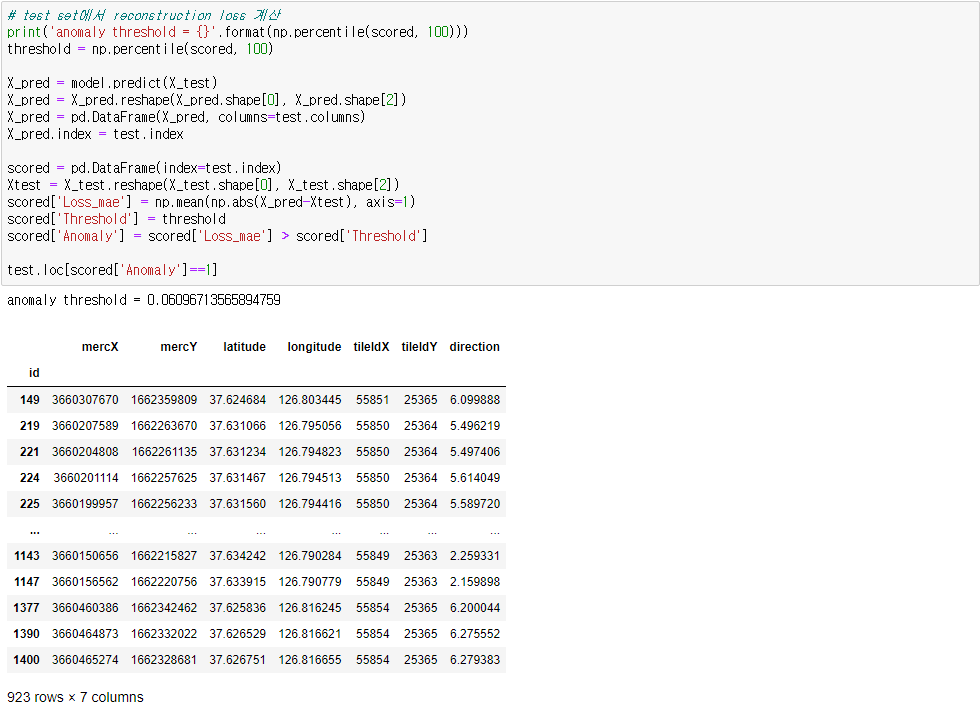


* + - **사용 변수**: log 관련 변수(mercX, mercY, …, direction)
    - **결과:**

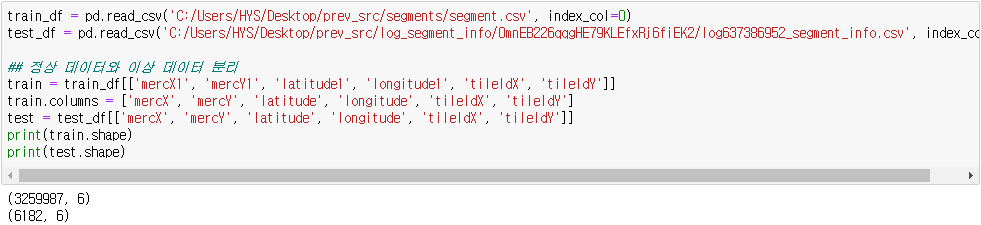




* + - * anomaly threshold = 0.06096713565894759



* + - * 923개의 이상치 발견
        + 학습 2와 개수는 비슷하지만 결과물은 전혀 다름
  + **학습 4**
    - **데이터셋**: segment.csv, log636556604\_segment\_info.csv
      * **training set**: segment.csv
      * **test set**: log636556604\_segment\_info.csv



* + - **사용 변수**: log 관련 변수(mercX, mercY, …, tileIdY) -> direction 제외
    - **결과: 진행중..**
* **학습 이후 문제점**
  + LSTM AE를 이용한 Anomaly detection은 Reconstruction error가 임계값을 넘어섰을 때 이상 데이터로 판단하는 것인데, 이상 데이터(잘못 매칭된 segment)를 찾아낼 뿐이지 새로운 segment를 찾지는 못하는 것 아닌가?
    - 예측 모델이 필요한 것이 아닌가?
    - **mercX, mercY, latitude, longitude, tileIdX, tileIdY를 이용해서 seg\_lat1, seg\_long1, set\_lat2, seg\_long2를 예측**
  + 데이터가 시계열 데이터긴 하지만, 이를 통해 찾아낸 이상치와 새로운 segment 가 무슨 연관이 있는가?
* [[코드리뷰]LSTM-based Anomaly Detection - 새내기 코드 여행 (joungheekim.github.io)](https://joungheekim.github.io/2020/11/14/code-review/)
  + LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection
* [LSTM Autoencoder for Anomaly Detection (velog.io)](https://velog.io/@jaehyeong/LSTM-Autoencoder-for-Anomaly-Detection), [LSTM AutoEncoder를 사용해서 희귀케이스 잡아내기 (tistory.com)](https://data-newbie.tistory.com/136)
  + curve shifting
* [LSTM AE를 이용한 시계열 데이터 이상 탐지 - (1) 개요 (velog.io)](https://velog.io/@jonghne/LSTM-AE%EB%A5%BC-%EC%9D%B4%EC%9A%A9%ED%95%9C-%EC%8B%9C%EA%B3%84%EC%97%B4-%EB%8D%B0%EC%9D%B4%ED%84%B0-%EC%9D%B4%EC%83%81-%ED%83%90%EC%A7%80-1-%EA%B0%9C%EC%9A%94)
  + 간단한 LSTM Autoencoder Anomaly detection
* [LSTM Autoencoder를 이용한 ECG 이상 탐지 (tistory.com)](https://quokkas.tistory.com/entry/LSTM-Autoencoder%EB%A5%BC-%EC%9D%B4%EC%9A%A9%ED%95%9C-ECG-%EC%9D%B4%EC%83%81-%ED%83%90%EC%A7%80)
  + LSTM Autoencoder for Anomaly Detection\_by Brent Larzalere
  + train data는 fit\_transform을, test dats는 transform을 사용하는 이유: [[scikit-learn] transform()과 fit\_transform()의 차이는 무엇일까? (tistory.com)](https://deepinsight.tistory.com/165)