2022.01.28

머신러닝
팀 프로젝트

제주도 내 퇴근시간 버스 승차인원 예측 모델 구축

데이터리안2
DATARIAN2

아시아경제 핀테크 디지털 금융사이언티스트 양성과정 팀 프로젝트 2

남경호 박광운

OUR ANALYSIS PROCESS

데이터 분석 과정

데이터리안2
DATARIAN2

문제 정의

데이콘 경진대회 소개
모델 평가척도 / 접근방식 소개
라이브러리와 사용 데이터 설명

첫번째

두번째

탐색적 데이터 분석(EDA)

데이터의 분포와 변수 간 관계 등
데이터의 특성 이해 및 인사이트 발견

데이터 전처리

머신러닝 모델 생성을 위한
데이터 전처리 및 변수 생성
및 생성된 전체 변수 정리

세번째

네번째

모델 구축 및 검증

최종 데이터 셋을 활용한
머신러닝 모델 구축 과정 및 예측 모델 검증
교차검증,변수선택, 그리드탐색

분석 한계점

데이터 분석 과정의
한계점

다섯번째

OUR PROBLEM

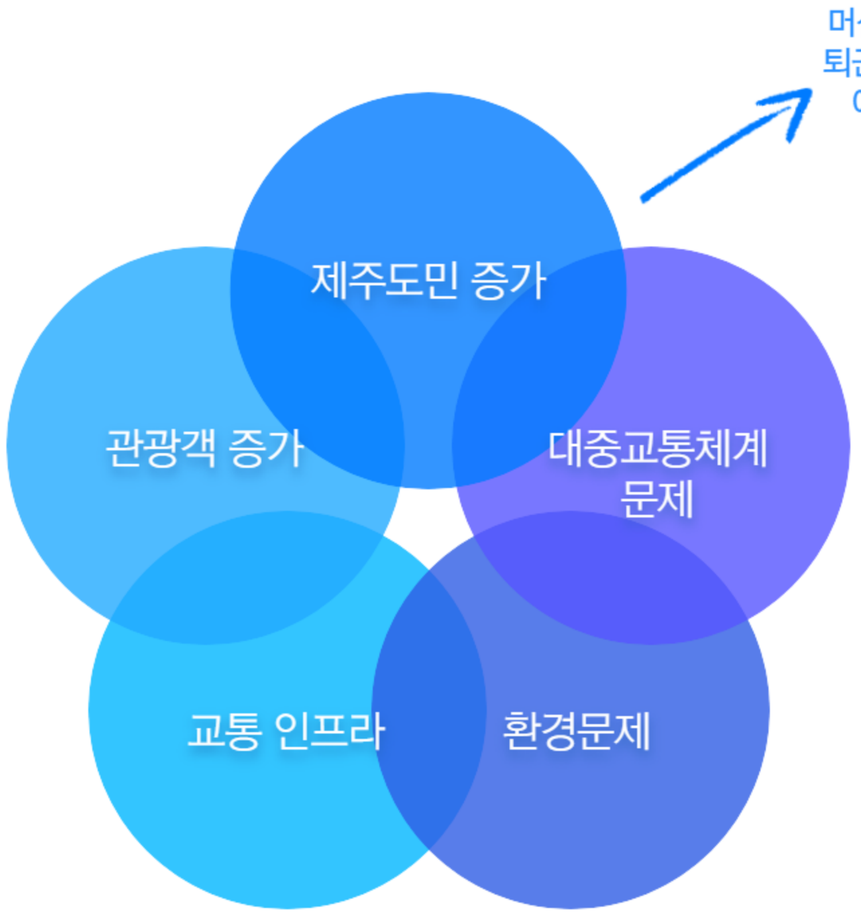
문제 정의

제주도의 교통체증 문제

제주도 내 등록인구는 19년도 11월 기준 69만명으로 **연평균 4%대 성장**을 보이고 있다.
또한 외국인과 관광객까지 고려하면 **상주인구는 90만명** 넘을 것이다.
제주도민의 증가와 외국인의 증가로 **제주도의 교통문제는 서울보다 심각한 상태**이다.
이 문제의 해결을 위해 데이콘에서는 제주도 공공 위치 데이터와 AI로 버스 승객 이용을 분석하고 교통문제를 해결하고자 **예측 경진대회**를 진행했다.

문제 해결을 위한 분석 목적

제주도 내 교통문제 해결을 위해 **제주도에서 운행중인 버스의 효율적인 운행이 필요**하다.
이를 위해 **제주도 버스의 퇴근시간 버스 승차인원을 예측하는 모델을 개발**하여
제주도 내 교통문제 해결의 도움을 주고자 한다.



제주도 내 교통문제

퇴근시간 승차인원 예측 | 모델 구축을 위한 EDA

데이터리안2 DATARIAN2

Dataset Splitting ▶ | TrainData는 9월의 데이터,
TestData는 10월의 데이터로써 출퇴근시간의 승하차 인원이 기록됨

Train Data / Test Data

A Number of Passengers ▶ | 시간대 별 승객 수 비교

Number of Rush hour Customers (18~20)▼|▲(8~10)

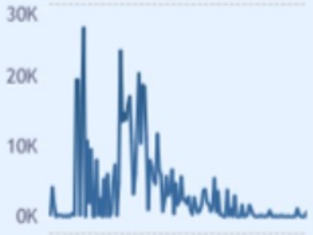
▶▶ 출근시간과 퇴근시간 승차인원이 비슷한 양상을 보임

Number of Rush hour Customers (18~20)▲|▼(8~10)

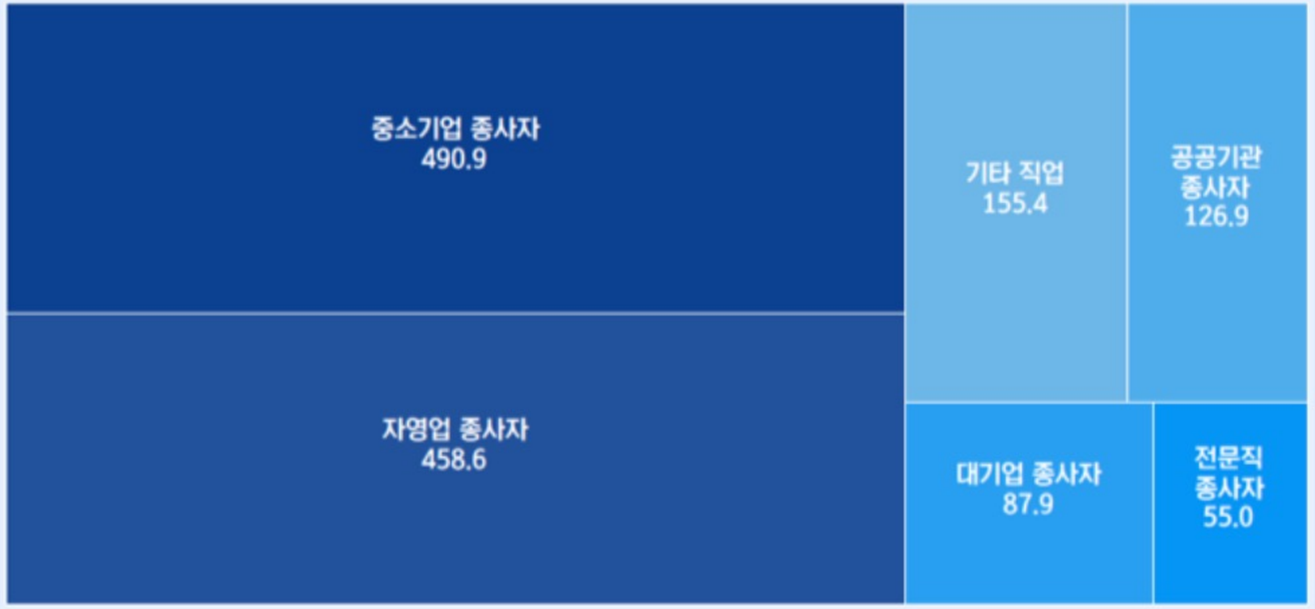
▶▶ 출근시간과 하차인원과 퇴근시간 승차인원이 비슷한 양상을 보임

number of passengers based on the interval between bus ▲|▼

▶▶ 배차간격이 짧은 버스는 퇴근시간 승차인원이 많음



Customer Segmentation ▶ | 각 직업군별 비율의 합계 트리맵



Weekday / Weekend / Area ▶ | 각 요일 별 데이터 수 비교,
주중/주말 데이터 수 비교, 제주시와 서귀포시 데이터 수 비교를 위한 그래프

요일별 탑승객 수 비교

주중 vs 주말 탑승객 수

제주시 vs 서귀포시 탑승객수

Domographic Information ▶ | 제주시 금융라이프 데이터로 보는 탑승객 상위 10개 동(읍) 직업군별 종사자 비율의 합계 (By Jeju Finance life Data)

Si	Dong	18~20 Ride	Job Majorc	Job Other	Job Profession	Job Public	Job Self	Job Smallc
제주시	이도이동	54,779	15.914	28.78	10.576	27.691	83.428	82.786
	연동	49,176	18.948	30.67	9.245	24.790	85.800	100.600
	노형동	45,178	19.919	34.40	12.723	26.357	85.365	94.557
	용담이동	36,428	6.245	10.29	3.477	8.725	33.048	40.787
	아라일동	27,246	5.317	11.40	4.493	9.341	22.364	23.839
	아라이동	20,941	1.574	4.66	1.763	3.029	11.817	10.769
	이도일동	19,071	5.157	9.05	3.482	7.041	30.003	31.405
	오라일동	16,901	1.883	3.35	1.044	2.891	12.722	13.826
	화북일동	15,998	6.867	12.04	3.578	9.397	40.402	47.512
	용담일동	8,221	3.359	5.80	2.395	4.295	19.474	23.653
서귀포시	서귀동	23,394	2.733	4.98	2.227	3.346	34.150	21.179

MODEL EVALUATION

모델 평가 개요

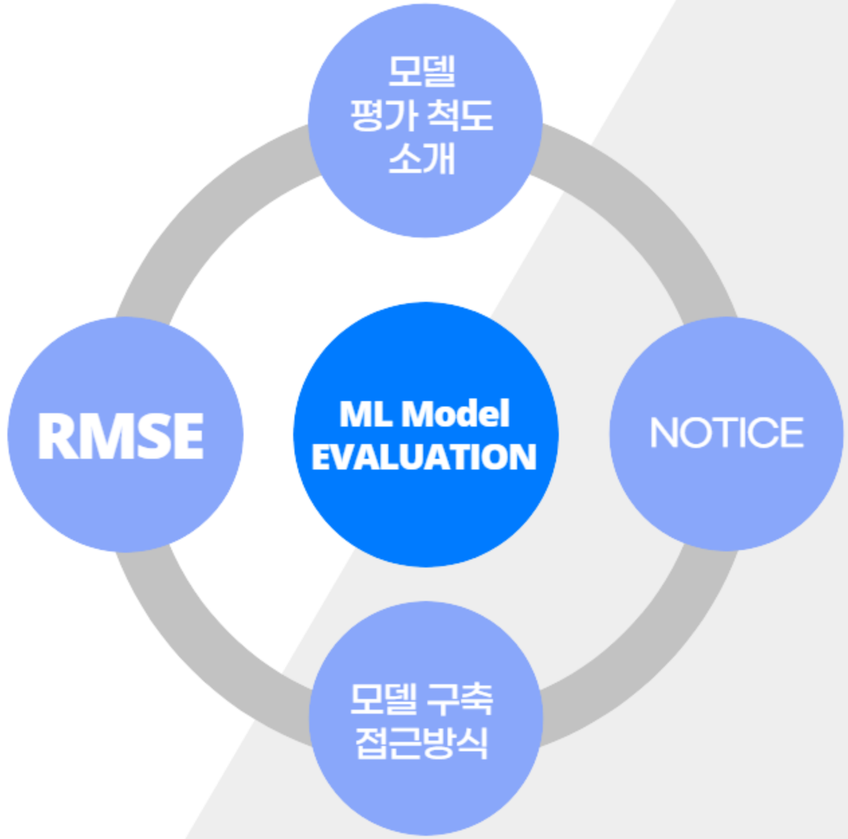
모델 평가 척도, RMSE

지도학습, 회귀(예측)모델

지도학습은 크게 분류와 회귀로 나뉨
본 모델 구축은 **승차인원을 예측**해야하므로
지도학습 중 회귀 문제에 해당함
회귀 모델을 평가하는 RMSE를 사용

RMSE

실제값과 예측값의 차이의 제곱 합을 n으로 나눈 뒤
제곱근을 구하면 모델 평가 척도인 *RMSE*
잘 학습된 모델일수록 RMSE가 낮음
RMSE를 최소화 하는 방향으로 모델을 만들어야 함



분석을 위한 접근방식

오전시간(06:00~12:00) 데이터 활용

오전시간(06:00~12:00)의 데이터만을 이용해
*퇴근시간(18:00~20:00)의 버스 승차인원을 예측*하는
모델을 만드는 것이 모델의 목적임.
따라서 데이터누수 방지를 위해 오전 데이터만 활용

9월 1일~30일 데이터 활용

주어진 학습 데이터 셋은 9월 1~30일 까지의 데이터
테스트 데이터 셋은 10월 1~15일 까지 데이터 이므로,
9월 한달 간의 데이터만 모델 학습에 활용
10월 데이터 사용시 이것도 데이터 누수에 해당

ANALYSTIC TOOL & DATA

분석 도구 및 데이터 소개

python

파이썬 라이브러리
numpy, pandas, seborn,
sciktlearn, geopy 등과
시각화 툴인 tableau를 활용

tableau

활용 도구

사용 데이터

내부 데이터

train data
test data
bus_bts (승객의 승하차 정보)
행정_법정동 중심좌표

외부 데이터

제주도 지역민 금융 생활 통계 자료
제주도 날씨, 기상정보 데이터

분석 도구인 파이썬 라이브러리와 시각화 툴인 tableau를 활용하여
데이콘에서 제공하는 데이터와 크롤링을 통하여 불러온 날씨 데이터를 분석하여
퇴근시간 버스 승차인원 예측 모델을 만들어보고자 함

INDEPNEDENT VARIABLE for Machine Learning

모델 구축을 위한 머신러닝 변수 생성

머신러닝 모델 구축을 위해 대회에서 제공해준 **내부 데이터를 통한 생성한 변수**들과
외부데이터인 **날씨 데이터, 제주도민의 금융생활 데이터**를 통해 변수들을 생성하고,
추가 변수 생성 및 라벨 및 원핫 인코딩을 통해 모델 구축을 위한 변수들을 생성해 주었다.

모델 구축을 위한 전체 변수

내부 데이터

요일, 요일별 평균 탑승객 수
버스 종류별 평균 탑승객 수
일별 오전 시간대 총 탑승객 수
배차간격, 연휴, 수요 예상 정류장

내부 데이터

승하차 시간대 통합 변수
오전 시간의 승객 수
카테고리 별 승객수의 합과 비율
인구밀집 지역, 버스정류장과의 거리

날씨 및 금융 데이터

버스 정류장과 날씨 측정소의 거리
지점별 기상정보 변수, 강수 여부
각 동별 금융 정보 관련(직업, 소득,
소비, 부동산) 관련 변수 등

추가 변수 및 라벨/원핫 인코딩

제주도 기상정보 변수
ID별 퇴근시간 총/평균 승객 수
라벨 인코딩(시내외 버스 여부,
주중/주말, 지리 관련 변수)

ML MODEL EVALUATION

머신러닝 모델 선택 및 검증

머신러닝 모델 선택

배깅 방식의 랜덤포레스트

랜덤포레스트 회귀모델을 선택한 이유는,
9월 한달 간의 데이터로 10월 중순까지의 데이터를 예측해야 하므로
짧은 시간의 학습 데이터로 일반화된 모델을 만들어야 하기 때문임.
일반적으로 부스팅계열 모델이 성능이 높지만, 부스팅계열의 모델은
오답에 가중치를 주는 방식으로 훈련하므로 학습 데이터의 기간이 짧다면
과적합 가능성이 높으므로 부스팅 계열 모델의 사용을 지양함.

퇴근시간 승차인원 예측을 위한
머신러닝 모델 선택

모델 생성 및 학습

배깅 방식의 RandomForest 모델 생성

모델 검증

교차검증 방법인 K-fold Cross Validation
· cv=5로 5-fold 검증 실시
· 5번 교차검증을 진행했으며 각각의 RMSE 값과
RMSE의 평균 값을 확인함

5번의 교차검증 결과 각각 RMSE 값은
2.37994074, 2.4957904, 2.52982728, 2.3599269, 2.24576673

RMSE의 평균 : 2.404

ML MODEL EVALUATION

변수선택과 하이퍼파라미터 튜닝

변수 선택 과정

01 랜덤포레스트 모델로 변수 A와 B를 기본 변수로 선택

02 기본변수 A와 B만으로 RF성능을 교차검증을 통해 RMSE 값 확인

03 RMSE값을 기준으로 변수를 하나씩 추가해 성능 확인

04 기준 RMSE보다 작을 때는 변수 추가, 클 때는 변수 제거 과정 반복

랜덤포레스트 모델로 A/B테스트 방식의 변수 선택 결과
총 113개의 변수가 선택되었으며, input_var1로 저장함

하이퍼파라미터 튜닝

GridSerchCV 그리드 탐색

탐색 과정 : n_estimaors는 [200, 300, 500]
max_features는 [5, 6, 8]
min_samples_leaf는 [1, 3, 5] 로 지정

cv=3, 변수는 input_var1으로 최적의 하이퍼파라미터 탐색
n_estimaors는 500, max_features는 8, min_samples_leaf는 1로 나오며,
세번의 교차검증의 결과 RMSE는 0.8692의 결과

RF모델 결과

최적의 하이퍼파라미터를 통해
랜덤 포레스트 값을 산출한 최종 RMSE 값은
0.8198

PROJECT BREAKING POINT

프로젝트 한계점 및 배운점

프로젝트 한계점

- 과적합 가능성이 높은 부스팅 계열의 모델을 지양하여 배깅방식의 랜덤포레스트만 사용했지만,
  예측을 위한 다양한 회귀모델을 사용하여 비교해보지 못한 점
- 프로젝트의 시간상 성능향상을 위한 다양한 단일 모델로 앙상블 모델을 시도해보지 못한 점이 아쉬움
- 다양한 변수선택법이 있음에도 A/B테스트 방식으로 하나씩 변수선택을 진행 하다보니
  변수선택에 많은 시간이 소요되었던 점

프로젝트 배운점

- 탐색적 데이터 분석부터 다양한 데이터 전처리, 모델 구축과 검증, 최적의 하이퍼파라미터 튜닝까지
  예측을 위한 회귀 모델 머신러닝 구축과정을 처음부터 끝까지 해볼 수 있었다는 점

감사합니다

2022.01.28

데이터리안2
DATARIAN2