Проект Автопати - Анализа на Безбедност на Делници

Оливер Бутески, 226023 - Даниел Костоски, 221189 - Виктор Христовски, 226026

Проектот е прикачен на следниот линк: https://github.com/viki123v/traffic-predictions

1 Подготовка на податоци

На почеток, податочното множество кое го добивме го исчистивме во добар csv формат.

Дополнително, го поделивме на 10 csv фајлови, каде има само една колона за просечен годишен дневен сообраќај и тежински индекс. Всуштност имаме 10 фајлови, еден за секоја година од 2014 до 2023.

Дополнително направивме уште еден фајл, каде последните две колони, за PGDS и Wi се просек од сите вредности од 2014 до 2023 за секоја делница.

2 Пресметка на важност на податоците кон тежинскиот индекс

За системот за пресметување на важноста на карактеристиките го користиме моделот Random Forest Regressor со комбинација на Permutation Importance техника за пресметувањето на самата важност, т.е. влијанието на податоците кон целната променлива.

Моделот е конфигуриран со 800 дрвја на одлучување. Тој работи со неограничена длабочина, додека минималниот број на примероци потребни за поделба на јазол е 5, а минималниот број на примероци во листовите е 2. За избор на карактеристики при секоја поделба се користи квадратен корен од вкупниот број на карактеристики.

Важноста се пресметува во повеќе чекори. Првично податоците се стандардизираат со Standard Scaler, потоа моделот се евалуира преку Repeated K-Fold крос валидација со 5 поделби повторени 5 пати.

По тренирањето на моделот на целиот датасет, преку permutation importance, за секоја карактеристика се применува следната процедура: вредностите на карактеристиката се случајно мешаат, се пресметува R^2 score со измешаните податоци, а падот во перформансите се третира како мерка на важноста на таа карактеристика. Процесот на мешање се повторува 20 пати за секоја карактеристика за да се добие стабилна проценка со пресметан просек и стандардна девијација.

Со тоа, резултатите на анализата се зачувуваат во CSV формат. За секоја година од 2014 до 2023, како и просекот, добиваме резултатна датотека каде што се дефинирани вредностите на важност на секој податок кон тежинскиот индекс.

3 Модел за пресметување на граница на безбедност

За оваа задача користиме кластеринг техники за да ги категоризираме патиштата според нивната безбедност брз основа на просечниот тежински индекс. Целта е да се креира модел што ги класифицира патиштата во IRAP категории (Green, Yellow, Orange, Red, Black) врз основа на историските податоци за сообраќајни несреќи.

Податоците ги земаме од претходно дефинираната датотека каде ги имаме просечните податоци за просечен дневен сообраќај и просечен тежински индекс на секоја делница.

Процесот на кластерирање се спроведува со неколку различни алгоритми за да се најде најдобрата сегментација на податоците. Се користат K-Means како центроиден метод и Agglomerative Clustering со различни linkage стратегии (ward, complete, average, single). Секој модел создава 5 кластери, што одговара на петте IRAP категории на безбедност. Пред кластерирање, податоците се стандардизираат со StandardScaler за да се обезбеди еднаква тежина на сите карактеристики.

Перформансите на различните кластеринг модели се евалуираат со три метрики: Silhouette Score (мери колку добро секој податок е сместен во својот кластер), Calinski-Harabasz Score (го мери односот на дисперзија меѓу кластерите и внатре во кластерите), и Davies-Bouldin Score (мери просечната сличност меѓу секој кластер и неговиот најсличен кластер). Резултатите покажуваат дека "complete"linkage методот дава најдобри перформанси со највисоки вредности за Silhouette и Calinski-Harabasz скорови.

По идентификацијата на оптималните кластери, се креира K-Nearest Neighbors (KNN) класификатор за да се предвидува припадноста на нови патишта. Моделот се тренира на 60% од податоците, валидира на 20%, и тестира на преостанатите 20%. Се врши fine-tuning на хиперпараметарот K (број на соседи) со тестирање на вредности од 3 до 11. Резултатите покажуваат дека K=3 дава конзистентно добри перформанси со F1 score, recall и accuracy од околу 90%.

Кластерите се мапираат кон IRAP категории врз основа на просечните вредности на индексот на несреќи во секој кластер. Кластерите се рангираат од најбезбедни до најопасни: Кластер 1 (просек 3.49) се мапира како Green (најбезбеден), Кластер 0 (11.49) како Yellow, Кластер 4 (22.20) како Orange, Кластер 3 (75.09) како Red, и Кластер 2 (105.68) како Black (најопасен). Ова мапирање овозможува директна интерпретација на резултатите во контекст на IRAP стандардите.

Финалниот модел се имплементира како WeightedAccidentsClassifier класа која комбинира pipeline од StandardScaler и KNN класификатор со речник за мапирање на кластери кон IRAP категории. Овој модел прифаќа просечна вредност на индекс на несреќи и враќа соодветна IRAP категорија (Green, Yellow, Orange, Red, или Black). Моделот се зачувува со joblib за понатамошна употреба, овозможувајќи брза и конзистентна класификација на патишта според нивната безбедност.

4 Модел за пресметување на тежински индекси

За оваа задача при вчитување на податоците, исто така ги користиме и податоците кои ги добивме од точка 1. - податоците за важноста на податоците кон тежинските индекси.

При тоа, спроведуваме повеќе тестови, каде правиме преглед, на кумулативно најважните 70%, 80% и 90% од податоците. За оваа цел, се вчитуваат дефинираните важности, се сортираат во опаѓачки редослед, и ги зачувуваме сите, се додека збирот на нивните проценти не го достигне дадениот праг.

Откако ги дефиниравме кои податоци ќе ги користиме во анализата, правиме тестирање на неколку регресиони модели. Тука се тестирани Linear Regression, MLP, XGBoost, LightGBM, Random Forest Regressor и Gradient Boosting. Сите овие се тестирани со комбинирање на сите три претходно дефинирани прагови. При тоа, секој модел се валидира користејќи 5 fold cross validation, со стандардизација на податоци и мерење на перформансите преку метриките RMSE, MAE и R^2.

Резултатите покажуваат дека најдобар модел е Gradient Boosting со 70% карактеристики, постигнувајќи RMSE од 6.97 и R^2 од 0.30, при што Gradient Boosting доминира во сите три тестирани сценарија.

На крај, најдобриот модел зачувуваме заедно со scaler-от и листата на карактеристики во joblib датотека, овозможувајќи лесен deployment на моделот за идни предвидувања.

5 Тестирање на моделите

За тестирање на моделите направивме FastAPI веб апликација која служи како API за предвидување на опасноста на дадени делници. За самата функција на апликацијата, тука ги користиме моделите кои ги направивме и зачувавме од точки 2. и 3.

Главната функционалност е имплементирана преку POST endpoint /predict кој прифаќа влезен стринг со 12 нумерички вредности одделени со точка-запирка, кои претставуваат карактеристики на патниот сегмент (должина, PCI, ограничување на брзина, број на мостови, просечен сообраќај, климатски услови, итн.) - Овие податоци ги дефиниравме во точка 3.

Овој стринг е парсиран и потоа податоците добиени се стандардизираат користејќи го зачуваниот scaler и се испраќаат до регресорот за предвидување на тежинскиот индекс. Отако тежинскиот индекс е дефиниран, се користи моделот за класификација на безбедноста и при тоа се добива една од IRAP категориите за безбедност.

Па апликација враќа JSON одговор кој го содржи предвидениот индекс на опасност и за тоа во која категорија спаѓа делницата. При тоа зависно во која класификација спаѓа, страницата има различен изглед, укажувајќи на предвидената IRAP категорија.