Предлог пројекта из СИАП-а

Овим документом дат је предлог пројекта из предмета Системи за истраживање и анализу података. Предочена је дефиниција проблема, као и мотивација. Након мотивације су наведени радови на сличну тему. Описан је скуп података који је укратко и објашњен, методологија решавања проблема, коришћен софтвер и, на крају, евалуација.

Дефиниција проблема

Предикција цене половних аутомобила на основу многобројних карактеристика аутомобила. Идеја је пружање цене аутомобила потенцијалним купцима или продавцима. Систем би на основу прикупљених и обрађених података требао да предвиди цену аутомобила. Сама цена доста варира од аутомобила до аутомобила, јер, чак, и исти модели аутомобила имају различиту опрему, годиште, пређену километражу, итд.

Мотивација

Овај систем би требао да олакша куповину као и продају аутомобила. Особе које желе да купе аутомобил, а, често, не знају како се крећу цене, моћи ће уз помоћ овог софтвера да унесу жељене карактеристике аутомобила, а систем ће им као повратну информацију дати цену аутомобила који задовољава наведене карактеристике. На овај начин се може избећи то да купац неки аутомобил плати изнад његове реалне тржишне цене. Такође, уколико продавац не зна која је реална цена његовог аутомобила, уносећи његове карактеристике, може да сазна колико новца би могао да добије за аутомобил.

Релевантна литература

- *Fahad Rahman Amik, Akash Lanard, Ahnaf Ismat and Sifat Momen* (2021) *Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh*

  https://www.mdpi.com/2078-2489/12/12/514

Тема рада: Предикција цене половних аутомобила у Бангладешу користећи технике машинског учења и проналажење битних карактеристика које утичу на предикцију цене аутомобила.

Методологије: У овом раду су коришћени различити алгоритми за регресију укључујући *Linear Regression, Lasso, Decision Tree, Random forest, Extreme Gradient Boosting*. Такође, коришћен је *GridSearchCV* за одређивање хиперпараметара сваког модела.

Подаци: Прикупљени су подаци са сајта bikroy.com коришћењем *Web Scraper*-а. Иницијални скуп података садржи нешто више од 1200 инстанци, и на основу 9 карактеристика се одређује цена. Карактеристике које се користе су бренд, модел,

годиште, километража, трансмисија, тип аутомобила, тип горива, назив и кубикажа мотора.

Евалуација: Подаци су подељени у односу 80:20, где је 80% података коришћено као тренинг скуп података, а 20% као тест скуп података. За евалуацију је коришћен широк спектар метрика, као што су R^2 *score, MSE, RMSE* и *MAE*.

Закључак: *Extreme Gradient Boosting* је дао најбоље резултате, овај модел је био у могућности да коректно предвиди цену у више од 91% случајева.

Коментар: Скуп података је релативно мали. Ми ћемо користити већи скуп података. Такође, методе коришћене за решавање овог проблема, биће коришћене и у нашем раду.

- *Sri Sai Ganesh Satyadeva Naidu Totakura Harika Kosuru* (2021) *Comparison of Supervised Learning Models for predicting prices of Used Cars*

https://www.diva-portal.org/smash/get/diva2:1609361/FULLTEXT02.pdf

Тема рада: Поређење различитих техника надгледаног учења за предикцију цене коришћених аутомобила.

Методологије: Поређене су следеће методе: *Decision Tree, Linear Regression, Random forest, Light Gradient Boosted Machine*.

Подаци: Скуп података је прикупљен са *craigslist.org*. Скуп података садржи преко 400 хиљада инстанци и цена се предвиђа на основу 11 карактеристика: регија, километража, годиште, тип горива, стање аутомобила, марка, број цилиндара, тип, боја, трансмисија, погон.

Евалуација: Подаци су подељени у односу 80:20. Коришћене метрике за евалуацију су R^2 и *Relative Error*.

Закључак: На основу R^2 метрике, која је најбоља метрика за одређивање мере тачности, *Light Gradient Boosted Machine* је имао најбољи резултат око 94% на тренинг скупу података. На тест подацима је резултат био око 91%.

Коментар: Иако је скуп података иницијално био велики (око 400 хиљада инстанци), избацивањем *null* вредности, скуп података је смањен на 1/8 оригиналног скупа података. Да ли је било могуће бољим руковањем *null* вредности смањити број избачених инстанци и задржати нешто већи обим скупа података?

- Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric (2019) *Car Price Prediction using Machine Learning Techniques*

https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf

Тема рада: Предвиђање цене аутомобила користећи технике машинског учења.

Методологије: Коришћене су три методологије: *SVM, Random forest* и *Artificial Neural Network.*

Подаци: Скуп података је скрепован са сајта autopijaca.ba. Иницијални обим скупа података је нешто више од 1100 инстанци, који је касније редукован на приближно 800 инстанци. За предикцију цене аутомобила узете су у обзир следеће карактеристике: бренд, модел, стање, тип горива, марка, снага у киловатима, трансмисија, километража, боја, град, држава, број врата и многи други (да ли има сензоре, климу, ...).

Евалуација: Тачност система износи око 92% (*accuracy*), такође, мерена је и грешка, али у раду није наведена која тачно врста грешке.

Закључак: У раду је прво израчуната тачност система коришћењем појединачних метода. Обзиром да су резултати били лоши (испод 50% тачности), аутори су дошли до закључка да појединачне методе не могу дати довољно добре резултате. Због тога су увели 3 категорије (*cheap, moderate, expensive*), при чему се користи *Random forest,* да одреди категорију аутомобила, а, затим, су за сваку категорију користили други метод рачунања тачне цене (*cheap – SVM*, *moderate – ANN*, *expensive - SVM*).

Коментар: Скуп података је мали. Потребно је прикупити већи скуп података, док је број атрибута на основу којих се врши предикција довољног обима.

Скуп података

Скуп података ће бити самостално креиран скрепованјем са сајтова https://www.olx.ba/ и https://www.polovniautomobili.com/. Цена ће се предвиђати на основу атрибута произвођач, модел, годиште, километража, тип горива, број врата, киловата, кубикажа, тип, коњских снага, погон, емисион стандард, величина фелги, трансмисија, број степени преноса, боја, музика, паркинг сензори, клима, светла и многа друга. Цена аутомобила је у опсегу од 25€-100.000€.

Методологије

За решавање проблема биће искоришћене следеће методе:

1. *Linear regression, Lasso, Ridge, Elastic Net*
2. *Random forest, SVM, Boosting algorithms*

Прикупљене податке ћемо претпроцесирати. Претпроцесирање података обухвата следеће:

- Рад са недостајућим вредностима (испробаћемо неколико техника, нпр. избацивање недостајућих вредности, замена са просечном вредношћу, линеарна регресија и сл.)
- Анализа *outlier*-а (избацивање, *winsorizing* и сл.)
- Анализа корелација између обележја (уклањање обележја која су у великој међусобној корелацији)
- Претварање текстуалних обележја у нумеричка
- Анализа дистрибуције вредности обележја
- Скалирање вредности (*minmax scaler*, z-score нормализација,…)

Будући да ћемо скуп података креирати *web-scraping*-ом, биће потребно ускладити вредности одређених обележја (нпр. уколико су цене у различитим валутама, свешћемо их на исту валуту).

Излаз из система ће бити предвиђена цена за дати аутомобил.

Напомена: Приликом израде пројекта, уколико пронађемо друге технике претпроцесирање података које би нам могле побољшати резултат, применићемо их.

Метод евалуације

Иницијално ћемо пробати унакрсну и обичну валидацију, поделом скупа података у односу 80:10:10. Обзиром да вршимо предикцију, користићемо метрику R^2, али ћемо користити и *RMSE* метрику. Евалуација се врши над тестним скупом података који ће бити издвојени на почетку и неће се користити приликом развоја система.

Софтвер

Апликација ће бити израђена у програмском језику *Python*. Програмски код за развој модела ће бити писан у *Jupyter*-у. Обзиром да ћемо имати веб апликацију, она ће бити развијена уз ослонац на *Vue.js* за фронтенд и *Python* + *Flask* за бекенд.

План

- Прикупљање података
- Анализа и трансформација атрибута и података
- Обучавање модела
- Евалуација система
- Израда веб апликације

Тим

Чланови тима су: Алекса Гољовић R2 29/2021, Јелена Цупаћ R2 30/2021, Милан Маринковић R2 31/2021.