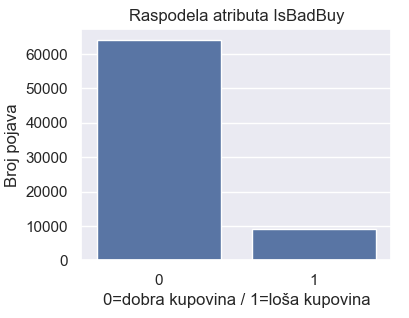
# Opis i razumevanje problema

Auto kuće kupuju mnoga polovna vozila na aukcijama sa ciljem da nabave što više automobila u najboljem mogućem stanju. Međutim suočavaju se sa rizikom kupovine vozila koji imaju nepredviđene probleme i ne mogu se dalje preprodati. Takva vozila takođe se mogu nazvati i “kiksevi”, odnosno promašaji. Ovi automobili mogu imati određene mehaničke ili probleme druge vrste, na primer “vraćene” kilometraže koji mogu sprečiti auto kuće da ih preprodaju, što rezultuje finansijskim gubitkom ovih ekonomski motivisanih subjekata. Glavni izazov ovog problema je predviđanje da li će vozilo biti “kiks” ili ne pre same kupovine i shodno tome minimizacija rizika auto kuća od loših investicija. Problem je posebno važan za auto dilere koji kupuju vozila na aukcijama. Rešenje bi im pomoglo da donesu bolje odluke pri kupovini i ponude svojim kupcima najbolji mogući izbor vozila.

# Opis i razumevanje podataka

Broj observacija u našem dataset-u iznosi 72 983, dok je broj atributa 34, odnosno 33 ulazne (input) varijable i 1 izlazna (output) varijabla. Izlazna varijabla je “**IsBadBuy**”. Ona pokazuje da li se automobil može smatrati kao “kiks”.

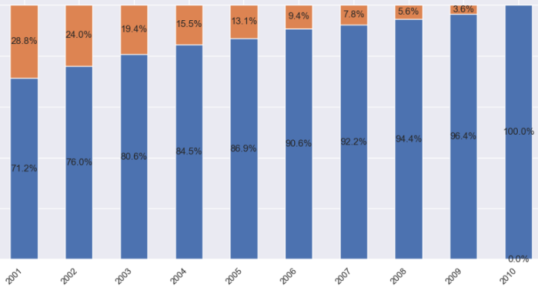
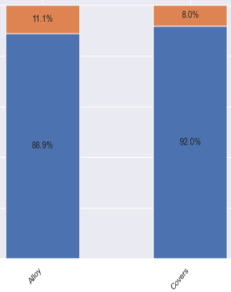
## Disbalans klasa

Na osnovu ovog grafika gde je prikazana raspodela izlazne varijable **IsBadBuy,** možemo primetiti da je klasa 0 (dobra kupovina) daleko zastupljenija od klase 1 (loša kupovina), tako da možemo zaključiti da postoji disbalans klasa.   
Ovaj disbalans može dovesti do kreiranja pristrasnih modela koji favorizuju većinsku klasu (dobra kupovina).

## Opis cene grešaka

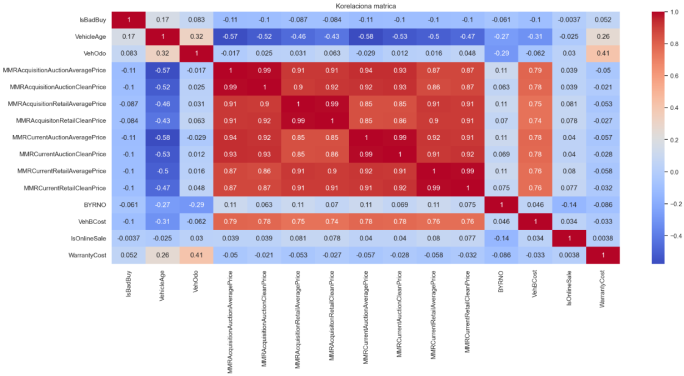
Ovo se odnosi na različite posledice pogrešnih predviđanja. U ovom kontekstu, postoje dve vrste grešaka: 1) False Positive – predviđanje da je automobil "dobar" kada je zapravo "kiks"; trošak kupovine problematičnog automobila (potencijalno veliki gubitak). 2) False Negative: Predviđanje da je automobil "kiks" kada je zapravo "dobar". Propuštena prilika za kupovinu dobrog automobila (izgubljena potencijalna zarada).

* 1. **Vizualizacija podataka**

Možemo uočiti trend da su manje šanse za kiksom što je vozilo novije (sl. 1). Zanimljiv podatak je da procenat kikseva veći za automobile koji imaju ratkapne u odnosu na one koji imaju felne od legure (sl. 2).

# Priprema podataka

Uklanjamo varijable koje nisu od značaja za predviđanje izlazne promenljive: **RefId** – ID vozila, **VehYear** – ima identično značenje kao VehicleAge, **WheelTypeId** – ima identično značenje kao VNZIP. Iz skupa podataka izbacujemo i varijable **AUCGUART** i **PRIMEUNIT** jer nedostajuće vrednosti čine oko 95%.

Iz korelacione matrice se može uočiti da postoje atributi koji imaju veoma visoku korelaciju (0.99). Zato su iz skupa podakata isključeni sledeći atributi: **MMRAcquisitionAuctionCleanPrice, MMR AcquisitonRetailCleanPrice, MMRCurrent AuctionCleanPrice** i **MMRCurrentRetail CleanPrice**.

## Nedostajuće vrednosti

Nedostajuće vrednosti su identifikovane u sledećim varijablama: Trim, SubModel, Color, Transmission, WheelType, Nationality, Size, TopThreeAmericanName, MMRAcquisition- AuctionAveragePrice, MMRAcquisitionRetailAveragePrice, MMRCurrentAuctionAverage- Price i MMRCurrentRetailAveragePrice. S obzirom da numerički atributi nemaju normalnu raspodelu, njihove nedostajuće vrednosti su zamenjene medijanom. NA vrednosti kategoričkih atributa su popunjene kategorijom sa najvećom frekvencijom. Izuzetak je atribut Nationality, čije NA vrednosti su popunjene ručno na osnovu kompanije koja proizvodi vozilo.

## Transformacija tipova podataka

U varijabli Trim su kategorije koje imaju manje od 500 observacija spojene u jednu kategoriju Other. Takođe, s obzirom da je za potrebe kreiranja određenih modela neophodno kategoričke atribute transformisati u numeričke, atributi Trim, Model, Make, SubModel i VNST su uklonjeni jer imaju prevelik broj kategorija za primenu get\_dummies metode.

# Treniranje algoritma

U skladu sa problemom koji je potrebno rešiti tj. binarnom klasifikacijom, odlučeno je da algoritmi za treniranje budu sledeći: Logistička regresija, Stablo odlučivanja, Naivni Bajes i KNN. Kako bi se poboljšali modeli skup podataka je izbalansiran nakon kreiranja prvih modela. Korišćeni su i ansambl algoritmi Gradient Boosting i Random Forest.

Vezano za konkretan problem, bitnije je identifikovati automobile koji predstavljaju promašaj pri kupovini. Zato je fokus na odzivu kao meri evaluacije modela. Pored odziva, mere kojima će modeli biti evaluirani su tačnost i AUC. Opravdanje za fokus na odzivu leži u tome što propuštena identifikacija automobila koji predstavlja kiks može dovesti do značajnih finansijskih gubitaka.

Na početku je naglašeno da je skup podataka neuravnotežen tj. da je procenat jedne klase u odnosu na drugu u izlaznoj varijabli 87.7:12.3 u korist „dobre“ kupovine. Ova disproporcija predstavlja uzrok visoke tačnosti i niskog odziva modela, što je i dobijeno kada su svi modeli trenirani na početnom nebalansiranom skupu podataka.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Nebalansirani skup | | | | Balansirani skup | | |
|  | **AUC** | **recall** | **accuracy** | **AUC** | **recall** | **accuracy** |
| NB | **0.66** | **0.24** | **0.82** | **0.65** | **0.53** | **0.61** |
| LR | **0.61** | **0.0** | **0.88** | **0.67** | **0.62** | **0.63** |
| THREE | **0.54** | **0.20** | **0.79** | **0.56** | **0.55** | **0.56** |
| KNN | **0.55** | **0.09** | **0.84** | **0.56** | **0.56** | **0.55** |
| Random Forest | **0.67** | **0.02** | **0.88** | **0.67** | **0.62** | **0.62** |
| Gradient boosting | **0.69** | **0.64** | **0.63** | **0.68** | **0.64** | **0.63** |

## Selekcija atributa

Za selekciju atributa korišćene su metode VarianceThreshold i SelectKBest. Prva metoda uklanja atribute čija je varijansa manja od 0.05, što je rezultovalo novim datasetom sa 26 atributa.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Balansirani skup | | | | Variance threshold | | |
|  | **AUC** | **recall** | **accuracy** | **AUC** | **recall** | **accuracy** |
| NB | **0.65** | **0.53** | **0.61** | **0.65** | **0.54** | **0.61** |
| LR | **0.67** | **0.62** | **0.63** | **0.67** | **0.61** | **0.62** |
| THREE | **0.56** | **0.55** | **0.56** | **0.56** | **0.55** | **0.56** |
| KNN | **0.56** | **0.56** | **0.55** | **0.57** | **0.57** | **0.56** |
| Random Forest | **0.67** | **0.62** | **0.62** | **0.67** | **0.62** | **0.62** |
| Gradient boosting | **0.68** | **0.64** | **0.63** | **0.68** | **0.63** | **0.63** |

Iz tabele vidimo da selekcija atributa VarianceThreshold metodom nije poboljšala performanse modela, čak se mogu uočiti i određena pogoršanja.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Balansirani skup | | | | Select K best | | |
|  | **AUC** | **recall** | **accuracy** | **AUC** | **recall** | **accuracy** |
| NB | **0.65** | **0.53** | **0.61** | **0.65** | **0.54** | **0.61** |
| LR | **0.67** | **0.62** | **0.63** | **0.67** | **0.62** | **0.62** |
| THREE | **0.56** | **0.55** | **0.56** | **0.56** | **0.55** | **0.56** |
| KNN | **0.56** | **0.56** | **0.55** | **0.57** | **0.57** | **0.56** |
| Random Forest | **0.67** | **0.62** | **0.62** | **0.67** | **0.62** | **0.62** |
| Gradient boosting | **0.68** | **0.64** | **0.63** | **0.67** | **0.66** | **0.62** |

Metoda SelectKBest je vrlo malo poboljšala performanse modela.

# Optimizacija hiperparametara

Nakon optimizacije, došlo je do značajnog poboljšanja performansi modela klasifikacionog stabla ali i većine drugih modela što se može videti u sledećoj tabeli:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Balansirani skup | | | | Optimizacija parametara | | |
|  | **AUC** | **recall** | **accuracy** | **AUC** | **recall** | **accuracy** |
| NB | **0.65** | **0.53** | **0.61** | **0.65** | **0.53** | **0.61** |
| LR | **0.67** | **0.62** | **0.63** | **0.64** | **0.60** | **0.60** |
| THREE | **0.56** | **0.55** | **0.56** | **0.66** | **0.62** | **0.61** |
| KNN | **0.56** | **0.56** | **0.55** | **0.62** | **0.57** | **0.59** |
| Random Forest | **0.67** | **0.62** | **0.62** | **0.68** | **0.64** | **0.63** |
| Gradient boosting | **0.68** | **0.64** | **0.63** | **0.68** | **0.64** | **0.63** |

# Zaključak

Korišćenje algoritama mašinskog učenja može značajno unaprediti poslovanje auto placeva i smanjiti šansu za gubitkom. Preduzeća koja prva primene ove metode u svom poslovanju razviće konkurentsku prednost i biti korak ispred ostalih na tržištu.

Algoritmi Random Forest, Gradient Boosting i Logistička regresija pokazali su se najefikasnijim u detekciji potencijalnih problematičnih vozila.

Zbog nedostatka vremena, nekoliko koraka koje bismo primenili su:

* Dublja analiza atributa kroz pronalaženje vrednosti k u SelectKBest metodi, kao i vrednosti za threshold u VarianceThreshold metodi koje bi nam obezbedile bolje performanse modela
* Primena drugih metoda selekcije atributa
* Pravljenje novih atributa čime bismo redukovali dimenzionalnost i stekli bolji uvid u podatke

*References*

[1] Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539-550

Adams, S., & Johnson, M. (2018). Comparative analysis of machine learning algorithms for predicting car auction outcomes. Journal of Artificial Intelligence Research, 25(4), 789-802.

Garcia, L., & Martinez, A. (2017). Impact of feature engineering on the performance of car quality prediction models. Data Science and Engineering Journal, 9(3), 210-225.

Thompson, P., & White, G. (2016). Ensemble learning strategies for improving predictive accuracy in used car quality assessment. Journal of Machine Learning Applications, 14(1), 45-60

Barua, S., Islam, M. M., Yao, X., & Murase, K. (2012). MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on knowledge and data engineering, 26(2), 405-425

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & electrical engineering, 40(1), 16-28.

Shamsudin, H., Yusof, U. K., Jayalakshmi, A., & Khalid, M. N. A. (2020). Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. 2020 IEEE 16th International Conference on Control & Automation (ICCA), 803-808

Fida, M. A. F. A., Ahmad, T., & Ntahobari, M. (2021, October). Variance threshold as early screening to Boruta feature selection for intrusion detection system. In 2021 13th International Conference on Information & Communication Technology and System (ICTS) (pp. 46-50). IEEE.