FRAUD DETECTION IN INSURANCE WITH MACHINE LEARNING MODEL

Dushko Todevski
UGD Shtip, Republic of North Macedonia, duskotodevski@gmail.com

Abstract: The subject of this paper is the presentation of a case of using artificial intelligence to increase the client base of companies. The data used in this model is from KAGGLE. This database consists of 1000 car accidents and car insurance claims from Ohio, Illinois and Indiana from January 1, 2015 to March 1, 2015. In this paper we present three models for artificial intelligence, logistic regression, gradient boosting and random forest. The results show that the best model can predict with 81% probability whether a potential customer would remain a customer of the company. Application of this model or similar models can be found in the detection of insurance fraud at the level of the entire industry, but also fraud at the level of certain insurance companies. This model is also a great potential for detecting tax evasion, but also in all other cases where there are similar situations and data available as in the presented case.
Keywords: fraud insurance, machine learning, artificial intelligence

ОТКРИВАЊЕ НА ИЗМАМИ ВО ОСИГУРУВАЊЕТО СО МОДЕЛ СО МАШИНСКО УЧЕЊЕ

Душко Тодевски
Универзитет „Гоце Делчев“, Штип, Република Северна Македонија
duskotodevski@gmail.com

Апстракт: Предмет на овој труд е презентирање на случај на користење на вештачка интлигенција за зголемување на клиентската база на компаниите. Податоците користени во овој модел се од KAGGLE. Оваа база на податоци се состои од 1000 автомобилски инциденти и побарувања за автомобилско осигурување од Охајо, Илиноис и Индијана од 01 јануари 2015 година до 01 март 2015 година. Во овој труд презентираме три модели за вештачка интелигенција, логистичка регресија, gradient boosting и random forest. Резултатите покажуваат дека кај најдобриот модел може со веројантост од 81 проценти да се предвиди дали еден потенцијален клиент би останал клиент на компнанијата . Апликативна примената на овој модел или слични модели можат да најдат при откривање на измами во сосигурањето на ниво на цела индусустрија, но и измами на ниво на одредени осигурителни компании. Овој модел исто така е голем потенцијал за откривање на даночни затајувања, но и кај сите други случаи каде што има слични ситуаии и податоци на располагање како во презентираниот случај.
Клучни зборови: измами осигурување, машинско учење, вештачка интелигенција

1. ВОВЕД

Една од главните трендови во сите индустрии па и во осигурувањето е работа во динамичен амбиент на дигитализација и со голема зависност од софтверски апликации и огромни бази на податоци. Во оваа смисла и современотото осигурување е под огромно влијание на овие процеси. Едната насока на влијание кај компанииите од осигурителната индустрија до дигиталните трендови овозможува креирање на решенија за подршка на осигурителните компаниите и потрошувачите во дигиталната економија, додека другата насока на влијание е во подршка на веќе постоечките осигурителни бизниси со нова технологија, софтвер, вештачка интелигенција и deep learning и во нивна комерцијалализација кај посточеките осигурителни бизнис процеси. Во суштина користењето на новите технологии е во голема мерка комплементарно со постоечките постулати на осигурувањето, особено во делот на комерцијализација на иновации како одговор на реалните потреби на бизнисот и потрошувачкото општество. Во овој трду конкретно ќе се задржам на користење на модел за машинско учење за откривање на осигурителни измами до делот на авто осигуривањето.

Овој труд е еден вид на надградба на еден од првите трудови на ова проблематика на (Bhowmik R, 2011) со наслов “Detecting auto insurance fraud by data mining techniques “ во кој се користат модели за машинско учење за предвидување на измамите во авто осигурувањето. Ова истражување треба да даде оценка дали со користење на понаков пристап на ова проблематика, пред се со користење на поинаков софтвер и на поинакви модели при предвидувањето може да се добие поголема предвидувачка моќ за дадениот пример.

Во овој труд се презентира еден вид на решение за подобрување на капацитетите за оценка на измамите во

осигуваурањето каде истото би се користело покрај веќе креираните процедури и постапки за контрола и превенција од измами. Истиот пример може да се користи и во развивање на слично решение за поддршка на случаи на измами или оперативен ризик во други индустрии. Овој модел со соодветна апликативна поддршка може да биде предмет на имплементнтација како сосфтверски модул во потоечките оперативни системи, но може да се користи и како додаток во некој од постоечките сервиси за продажба и односи со клиенит кои ги користат бизнисите. Потеницијалот на овој случај и решение на коиристење на вештачка интелигеница во продажбата е и во можноста во целост да се понуди како веб сервис на осигурителните компании кои секојдневно се среднуваат со побарувања на наплата на штети. Со ова решение на осигурителните компании дополнително ќе им се укаже на високиот ризик кај одреден побарувања на штети, со што ќе можата да посветат повеќе ресурси и внимание при наплатата на штета од одредени клиенти.
Основна цел на ова истражување е да предложи можност на користење на најновите модели на вештачка интелигенија за целите откривање на измамите при наплатата на штети во осигурувањето. Во ова истражување се презентирааат податоци од американската осигурителна индустрија, поточно од три држави. Основната идеја и централна хипотеза на ова истражување е да се испита можноста од креирање на модел кој би можел да предвиди дали одреден одреден пријавена штета е измама и која е веројатноста истата да е измама според моделот на машинско учење. Во случајов се користат податоците од од 1000 случаи на штети во авто осигурувањето од кои за дел е известено дека се измама, додека друг дел се регуларно пријавени и процесирани како штети. За потребите на ова истражување се користат 37 варијабли.

2. ПОДАТОЦИ И МЕТОДОЛОГИЈА

Податоците користени во овој модел се од KAGGLE. Оваа база на податоци се состои од 1000 автомобилски инциденти и побарувања за автомобилско осигурување од Охајо, Илиноис и Индијана од 01 јануари 2015 година до 01 март 2015 година. По опсежно истражување ова беше една од ретките целосно достапни базаи за истражување и аплицирање на ваков тип на модел.
Овој сет на податовци во има вкупно 39 варијабли. Во оваа база на податоци нам достапен податок дали податоците се од повеќе осигурителни компании или само од една компанија.
Некои од податоците кои се користет вклучува информации за:
Детектирана измама како таргетирана варијабла, додека другите варијабли опфаќаат демографки податоци и детали за осигурителната полиса и настанатата штета
Демографски информации за клиенти - пол, возрасен опсег и ако тие имаат партнери и зависни лица, но и колкав работн стаж имаат
Влезни варијабли:

1. Месеци_ како_корисник
2. Возраст
3. Број на полиса
4. Датум_поврзаност на датумот
5. Држава_политика
6. Полиса покриеност
7. Полиса _повратна
8. Полиса _годишен_премиум
9. Франшиза_ограничување
10. Осигуреник_реонски код
11. Осигуреник_пол
12. Осигурено_на ниво на образование
13. Осигурена_ занимање
14. Осигурено_ хоби
15. Осигурена_работа
16. Капитална добивка
17. Загуба на капитал
18. Датум_инцидент
19. Тип на инцидент
20. Тип на судир
21. Сериозноста на инцидентот
22. Власти_контактирани
23. Инцидент-држава
24. Инцидент-град

25. Инцидент_локација
26. Инцидент на час
27. Вклучен_ број_ на_во возила
28. Штета на имотот
29. Телесни_повреди
30. Сведоци
31. Полиција_извештај_достапност
32. Тотална_побарување_висина
33. Повреда_побарување
34. Побарување за имотот
35. Побарување од возило
36. Автоматско создавање
37. Автомобил_модел
38. Автомобил-година
39. Детектирана измама

За потребите на предвидувањето на овој модел се користат пред селектирани три модели на машинско учење и тоа логистичка регресија, random forest и gradient boosting. Примарно беа тестирани повеќе модели достапни во Orange data mining софтверот (Demšar et al., 2013; Demšar and Zupan, 2013), но при тестирањето со почетните параметри на модели претходно споменатите три модели се издвојуваа со резултатите на тренинг и тест податоците.

Логистичката регресија е класификационен алгоритам со примена на логистичка регресија и во овој случај користиме ласо регресија како параметар за класификација на независните ваирјабли, сила на cost функцијата од C=0.500. Random forest (Случајна шума) е комплексен метод на учење кој се развива со надградба не некој поедноставен модел најчесто моделот на дрва на одлучување. Овој модел најчесто се користи за класификација, регресија и други задачи. Прво беше предложен од Тин Кам Хо и понатаму го развија Лео Брејман (Брејман, 2001) и Адел Катлер. Овде ќе се користи моделот од Orange верзијата на софтверот (Demšar et al., 2013; Demšar and Zupan, 2013). Random forest (Случајна шума) генерално се базиран на градење на пакет дрва на одлучување. Секое дрво е развиено од примерок за подигање од трениранитe податоци. При развивање на одделни дрвја, се извлекува произволна подгрупа на атрибути (оттука и терминот „Случаен“), од кој е избран најдобриот атрибут за поделбата. Конечниот модел се заснова на мнозинството гласови од индивидуално развиените дрвја во шумата. Како параметри овде користиме 100 дрва и ограничување на под примероците да не бидат помали од 50 единици.

Gradient boosting е техника на машинско учење креирана да оговори на проблеми поврзани со регресија и класификација. Со овој алгоритам се генерира модел на предвидување кој е надградба на поедноставни модели на предвидување, обично дрва на одлуки (Friedman, 1999). Параметри кое се користеа во овој модел се: броеви на дрва 500, стапка на учење 0,095 со репликативен тренинг. За контрола на растот се користеа лимити од 5 на индивидуални дрва и ограничување на подпримероци со прараметар не помали од 3. Делот на тренинг истанциите беше сетиран на 0.85. а пресметки се користеше Gradient boosting од Scikit learn библиотеката

За потребите на предвидувањето на за останување во деловен одоно со телекомот ќе се користат повеќе техники за конструкција, тренирање и тестирање на моделите.

За потребите на крерање на примероци за тренирање и тестирање од почетните податоци за корисниците 1000 индивидуални инциденти, со што се прави поделба на почетниот примерок на два посебни дата сетови и тоа на тренинг и тест дата сет. Поделбата е направена во 80 проценти во корист на тренирањето на податоците, додека 20 проценти од податоците ќе се користат кај тест податоци за да се види успешноста на моделот со податоци кои не биле претходно користени за тренирање на моделите.

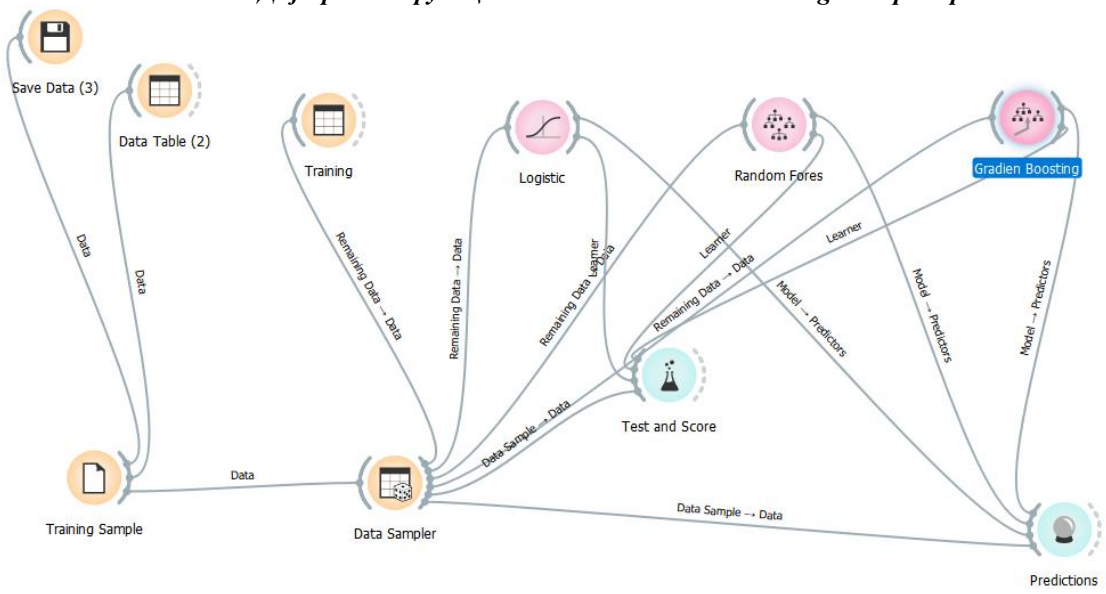
Фазата на тренирање на моделите се врши на тренинг податоците и тие по правило се секогаш во што е поголем обем за моделите да можат што е можно подобро да ги опфатат скриените релации и шаблоните кои настанале помеѓу вредностите на варијаблите. Ова практично значи дека моделите секој според капацитетот и перформансот на својот алгоритам се обидува да ја пронајде релацијата помеѓу триесет и осумте варијабли во однос податоците од нивните извештаите за пронајдена измама.

Logistic regression, random forest и gradient boosting моделите се користет за предвидување на податоците на у таргетирана променливата и кај трениранитe и кај тестираните податоци при што се добиваат различни оценки за секој од моделите при тренирање и тестирањет на податоците.

Оценки на моделите ќе се врши во две фази од кои во едната фаза ќе се врши тестирање само на трениранитe модели со претходно трениранитe подататоци(примерок од 80%) со стратифицирана крос-валидација на 10

различни примери. Другата фаза на финално оценување ќе се реализира со оценување на моќта на предвидување на моделите на 20 проценити од податоците кои не се користеа за тренирање и се одвоија од првичниот примерок со цел да се провери прецизноста на конечно тренираните податоци.

Слика 1 Дијаграм на функци на моделот- извод од Orange3 софтверот



Оваа слика ги прикажува визуелно сите мапирани процеси од реализација на истражувањето од фазата на користење на податоците, предпроцесирање, креирање на тренинг и тест примерок, па се до користење на моделите во тест и финалната фаза.

3. РЕЗУЛТАТИ И ДИСКУСИЈА

По спореведената конструкција на моделот и сите неопходни релации кои беа предходно дефинирани се чекори пред фазата на интерпретација на разултатите. При оценувањето на двете фази користење на моделите се користат следните параметри за оценување на спецификацијата на моделот:

- AUC (простор под ROC[5])
- CA (classification accuracy) е пропорција на точно класифицирани инстанци
- F1 е пондерирана хармонична средина на прецизност и recall параметарот
- Precision (прецизноста) е процентот на вистински позитиви предвидувања меѓу инстанците класифицирани како позитивни
- Recall процентот на вистински позитиви меѓу сите позитивни примери во податоците, на пр. бројот на болни кај сите дијагностицирани како болни.

За тест фазата на оценки на предвидувањата се користи примерок од податоците за 800 инциденти, и предвидувањата ќе се вршат со моделите тренирани на овој обем на податоци.

[5] ROC крива е график што ја прикажува работата на моделот на класификација на сите прагови на класификација. Оваа крива исцртува два параметри вистинска позитивна стапка на предвидувања и лажна позитивна стапка на предвидувања

Табела 1 Оценки од тест фаза на предвидувања

Модел	AUC	CA	F1	Precision	Recall	Просек AUC/CA
Random Forest	0.821	0.735	0.632	0.806	0.735	0.778
Logistic Regression	0.820	0.790	0.783	0.781	0.790	0.805
Gradient Boosting	0.856	0.790	0.781	0.779	0.790	0.823

Согласно презентираното највисоки оцени кај моделите во финалната фаза на оценка на моделот највисока прецизност има кај Gradient Boosting моделот со 85,6% прецизност според AUC и 79% според CA, или во просек прецизносност во однос на двата параметри од 82,3%. Просекот кај Logistic Regression е 80,5% додека кај Random Forest моделот е 77,8%. Овој резултат значи дека овој модел може да ја предвиди вредноста на таргетираната у (во случајот одредена пријава на штета е измама или не со 82,3%, 80,5% и 77,8 % кај Gradient Boosting, Logistic Regression и Random Forest моделот.

За финалната фазата оценки на предвидувањата се користи претходно одвоениот примерок кој ги содржи податоците за 200 инциденти (20% како тест примерок, додека 80% или 800 инстатнци се користеа за тренирање на моделот). Мора да се напомени дека предвидувањата и во тест и во финалната фаза се вршат со моделите тренирани на 80% (дата сетот од 800 инстанци) од тренинг примерок на податоци.

Табела 2 Оценки од финалната фаза на предвидувања со тест податоци

Модел	AUC	CA	F1	Precision	Recall	Просек AUC/CA
Random Forest	0.832	0.710	0.639	0.637	0.710	0.771
Logistic Regression	0.839	0.775	0.767	0.764	0.764	0.807
Gradient Boosting	0.821	0.795	0.794	0.794	0.795	0.808

Согласно презентираното највисоки оцени кај моделите во финалната фаза на оценка на моделот највисока прецизност има кај Gradient Boosting моделот со 82,1% прецизност според AUC и 79,5% според CA, или во просек прецизносност во однос на двата параметри од 80,8%. Просекот кај Logistic Regression е 80,7% додека кај Random Forest моделот е 77,1%. Овој резултат значи дека овој модел може да ја предвиди вредноста на таргетираната у (во случајот одредена пријава на штета е измама или не со 80,8%, 80,7% и 77,1 % кај Gradient Boosting, Logistic Regression и Random Forest моделот.

Во случајот на финалното оценување имаме значително подоби резултати од оценувањето со трениранитe податоците од тест фазата, и може да се забележи дека Gradient boosting моделот во финалната фаза на оценување на моделот го завзема првото место за многу мала разлика, додека таа разлика беше многу поголем при тест фазата на оценување.

Значајно во двата случаи е што трениранитe модели можат без никаков проблем да се користат за апликација на овој тип на случаи и со овој тип на податоци. Препорака за имплементација на овие модел е дека прецизноста од во просек од 80-81% кај gradient boosting и logistic regression моделите може да има големо значење за оценување на одреден инцидент дали е измама или не. Овие модели даваат силна препорака дека вака конструирните модели можат со голема веројатност да предвидат кои од пријавнеите штеи имаааат голем веројатност подоцна низ севкупна валидизација да се потврдат како имзами . Ова практично значи дека осигурителните компнии можата да се фокусираат на одредени случаи за кои моделот со голема веројатност ќе укаже дека се работи за измама.

4. ЗАКЛУЧОК

Ова истражување користи модели на вештачка инелигенција и машинско учење за предвидување на измамите на клиентите при поднесувањена штета кај авто осигурениците кај дел од анонимни осигурителни компании во САД. Идејата на овој труд е преку креирање на модел кој ќе произлезе од постоечките достапни податоци поврзани со инциденти за настанати штети, да се истражи можноста и потенцијалот за предвидување и класифицирање на доставените идни барања за штети како потенцијални измами или не.

Во овој труд се користат податоци од анонимни осигуретлни компании во САД и нивни инциденти кои се класифицирани како измами или регуларни штети. Главанта идеја е да се испита предвидувачкиот капацитет на овој дата примерок со помош на Orange 3 софтверот 3.29.3 и да се согледа предвидувачкиот капацитет на

овие модели и тренинг податоци.

Резултатите на оценуваните модели во тест и во финалната фаза покажуваат прецизност од околу 81 проценти во просек кај Gradient boosting и Logistic regression моделите кој би бил најсоодветен за имплементација во некое решение го предлагаат gradient boosting моделот со просечна прецизност од 94,5 проценти.

Овој резултат и ова ниво на прецизност покажува дека доколку би имале податоци како во првичниот примерок, со 81 проценти прецизност би можеле да оцениме дали дали одреден поднесен инцидент за штета би бил измама или е регуларен случај на поднесување на штета. Со ова практично можеме да предвидиме дали еден инцидент ие потенциjална измама или не и со ова да се сугерира на служните да посветат повеќе внимане на овие предмети за да не се случи измама. Со подетална анализа на факторите на овој модел би можеле и да се добијат првични сознанија за главните фактори кои придонесуват за да еден инцидент се пријавуа како измама, поточно кли се главните параметри кои можата уште во рана фаза да укажат на детекција на потенцијални измами и така да се конструира систем за превенција од овој тип на ризик.

Апликативна примената овој модел или слични модели можат да најдат при детектирање на измами во многу индустрии и јавни служби но секако е потребно да има достапна база на служаи кои досека се откриени како измама во текот на работењето на деловниот субјект. Секако истио модел со помали или поголеми модификации може да се аплицира кај секоја компанија која има малку поголема база на податоци и особено е погодна во случи каде корисниците треба да добијат државни или приватни средства ако резултат на некоја деловна трансакција или придобивка.

БИБЛИОГРАФИЈА

Bhowmik R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*, 2 (4), pp.1–6. [Online]. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.3602&rep=rep1&type=pdf [Accessed 6 July 2021].

Bhowmik R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*, 2 (4), pp.1–6. [Online]. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.302.3602&rep=rep1&type=pdf [Accessed 6 July 2021].

Demšar, J. and Zupan, B. (2013). Orange: Data Mining Fruitful and Fun-A Historical Perspective. *Informatica*, 37. [Online]. Available at: http://oranga.biolab.si [Accessed 27 May 2021].

Demšar, J. et al. (2013). Orange: Data Mining Toolbox in Python Tomaž Curk Matija Polajnar Laň Zagar. *Journal of Machine Learning Research*, 14. [Online]. Available at: https://www.jmlr.org/papers/volume14/demsar13a/demsar13a.pdf [Accessed 27 May 2021].

Dhieb N et al. (2019). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In: *IEEE International Conference on Vehicular Electronics and Safety (ICVES)*. 2019. pp.1–5. [Online]. Available at: doi:10.1109/ICVES.2019.8906396 [Accessed 6 July 2021].

Dhieb N et al. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. *IEEE Access 8*, pp.58546–58558. [Online]. Available at: https://ieeexplore.ieee.org/abstract/document/9046765/ [Accessed 6 July 2021].

Dua P and Bais S. (2014). Supervised learning methods for fraud detection in healthcare insurance. *In Machine learning in healthcare informatics, Springer, Berlin, Heidelberg*, pp.261–285. [Online]. Available at: doi:10.1007/978-3-642-40017-9_12 [Accessed 6 July 2021].

Friedman, J. H. (1999). Stochastic Gradient Boosting. *Citeseer*. [Online]. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.1703&rep=rep1&type=pdf [Accessed 27 May 2021].

Itri B et al. (2019). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In: *Third International Conference on Intelligent Computing in Data Sciences (ICDS)* . 2019. pp.1–4. [Online]. Available at: https://ieeexplore.ieee.org/abstract/document/8942277/ [Accessed 6 July 2021].

Kirlidog M and Asuk C. (2012). A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences*, pp.989–994. [Online]. Available at: https://www.sciencedirect.com/science/article/pii/S1877042812036099 [Accessed 6 July 2021].

Roy R and George KT. (2017). Detecting insurance claims fraud using machine learning techniques. In: *International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. 2017. pp.1–6. [Online]. Available at: https://ieeexplore.ieee.org/abstract/document/8074258/ [Accessed 6 July 2021].