MEF UNIVERSITY

SALES LEAD RANKING FOR EXTENDED COVERAGE AUTOMOBILE INSURANCE POLICIES USING THE ONLINE QUOTES

Capstone Project

Ahmetcan Tekince

İSTANBUL, 2018

MEF UNIVERSITY

SALES LEAD RANKING FOR EXTENDED COVERAGE AUTOMOBILE INSURANCE POLICIES USING THE ONLINE QUOTES

Capstone Project

Ahmetcan Tekince

Advisor: Prof. Semra Ağralı

İSTANBUL, 2018

MEF UNIVERSITY

Name of the project: Sales Lead Ranking For Extended Coverage Automobile Insurance Policies Using the Online Quotes
Name/Last Name of the Student: Ahmetcan Tekince
Date of Thesis Defense: 27/12/2018

I hereby state that the graduation project prepared by Ahmetcan Tekince has been completed under my supervision. I accept this work as a "Graduation Project".

27/12/2018
Prof. Semra Ağralı

I hereby state that I have examined this graduation project by Ahmetcan Tekince which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

27/12/2018

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of Ahmetcan Tekince and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof. Semra Ağralı	………………………..
2. Prof. Özgür Özlük	………………………..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Ahmetcan Tekince

Name Date Signature

EXECUTIVE SUMMARY

SALES LEAD RANKING FOR EXTENDED COVERAGE AUTOMOBILE INSURANCE POLICIES USING THE ONLINE QUOTES

Ahmetcan Tekince

Prof. Semra Ağralı

DECEMBER, 2018, 28 pages

This study analyzes the features that are important for the extended coverage automobile insurance sale decision of the client and the improvement strategies for insurance sales using the information gained from the analysis of algorithms. We start with a binary classification stating that whether a sale is made after each quote or not. All quotes are scored and ranked in the decreasing order in which a sale was predicted but not realized. We use the Two-Class-Boosted Decision Tree, Two -Class Neural Networks and the Two-Class Locally Deep SVM models. The Neural Network model provided the best results; and a list of quotes that were not sold and also seemed very possible to be converted into sales was generated, which can be used by the sales staff for realizing these sales.

Key Words: Sales lead ranking, casco, automobile insurance, online insurance quotes, decision trees.

ÖZET

ONLINE VERİLEN KASKO TEKLİFLERİNDEN POTANSİYEL MÜŞTERİ SIRALAMASI

Ahmetcan Tekince

Tez Danışmanı: Prof. Dr. Semra Ağralı

ARALIK, 2018, 28 sayfa

Çalışmamızda müşterilerin verilen teklifleri satın alma kararını etkileyen nitelikleri tespit etmeye çalışıyoruz. Ayrıca bu bilgi ile satışların nasıl artırabileceğini de inceliyoruz. Çalışmamız esasen bir ikili tasnif çalışmasıdır, verilen teklifin satışa dönüp dönmediğini belirtir. Bununla elimizdeki teklifler içinden satış olarak tahmin edilen ancak satılmamış tüm teklifleri puanladık. Bu liste, satış çalışmalarımızda tercih edilecek kısa liste olacak. Kullandığımız modeller, ikili tasnif karar ağacı, sinirsel ağlar ve ikili yerel derin SVM olmuştur. Modeller arasından, Sinir Ağları modeli en iyi sonuçları verdi. Satış personeli tarafından kullanılabilecek, satılmamış ve satışa dönüştürülmesi çok mümkün görünen tekliflerin bir listesi hazırlandı.

Anahtar Kelimeler: Potansiyel müşteri listesi sıralaması, kasko, online sigorta teklifleri, karar ağaçları, SVM.

TABLE OF CONTENTS

1. INTRODUCTION

This study aims to discover how a client decides to accept or not an automobile insurance (extensive coverage for car insurance) quote he receives online. We use the data obtained from the insurance aggregator website "koalay.com." Koalay is the only online insurance aggregator and provider in Turkey that can finalize all processes from data input to payment and publish the policy as a pdf file. No agent involvement is required in the whole process. For this project we have access to all online sales data.

Koalay is the online brand of Telesure Sigorta. Telesure Sigorta ve Reasurans Brokerligi was established in Istanbul in 2013. Koalay is a registered and licensed insurance & reinsurance brokerage firm, and they are a part of the Guernsey-based international financial services group, BHL Holdings Limited (BHL). Koalay is a pioneer in a competitive industry, and they are determined to change the way people buy and shop for insurance. They aim to assist customers with their search for car insurance products by offering them multiple insurance quotes and making a recommendation for the best price that will fit their specific needs. Their approach is convenient and simple. Customers can instantly purchase or renew their insurance online or via phone (Telesure, 2018) (Koalay, 2018)

Since koalay.com is an online business and process a high amount of offers, we have access to an extensive amount of data. Due to the nature of online businesses our data is imbalanced. Generally, more than 98% of the quotes given to customers are not accepted. The sales ratio is around 1.6%. In order to overcome the issues that an imbalanced dataset brings, we searched the literature on processing imbalanced data.

As Weiwei Lin and Longxin Lin summarized very well in their 2017 dated study (Weiwei and Lin, 2017):

- "The traditional marketing method of selling insurance is mainly based on off-line sales business. Insurance salesmen sell the company's products by calling or visiting the customers. This blind marketing way has achieved good results in the past, which maintained the company sales performance for a long time through widespread

> sales. With the gradual opening of the insurance industry, a large number of private insurance companies enter the market, which forms a healthy competitive environment and constantly promote the reform of the insurance industry. On the other hand, people's willingness to purchase insurance gradually increased, the potential insurance customers are rapidly expanding. According to statistics, the success rate of the traditional telephone sale is less than one thousandth, and the insurance sales rate of a senior insurance salesmen can reach about two percent, but this is obviously very inefficient. Therefore, how to better accurately understand the users' purchase intention has become a very urgent need for the insurance company."

Shortly we used SMOTE to balance the data. Our data set was as clean as possible since it was our online sales data generated by our native system. However, there are still some flaws because of the nature of the business, which dictates using insurers' API data as well. So, there was some cleaning need to be done. The major one was the lack of Vehicle Value on some quotes. This is very important for us even before the analysis for prediction, so all quotes that do not have the Vehicle Value were deleted.

After we clean the data, we first chose the Two-Class Decision Tree model for predicting the false positive we aimed for. The Locally Deep SVM was our second model.

Since Azure ML provided a significant CPU power for us, we also wanted to try the Two-Class Neural Network model. Compared to the others, model training took so long for the Neural Network algorithm. However, it was worth it since prediction results were much better with Neural Networks model.

2. LITERATURE REVIEW

2.1. Imbalanced Data Literature Review

Imbalanced data is a common problem in data analytics, not just in online sales. Medical imaging and face recognition applications have much higher risks when predictions fail. We found that SMOTE (Synthetic Minority Oversampling Technique) as Blagus and Lusa (2013) stated; is usually the appropriate way of processing the imbalanced data. (Blagus and Lusa,

2013) After data cleaning and missing data issues are solved, SMOTE can be applied. (Fernandez, Garcia, Herrera, Chawla, 2018) (SMOTE, 2018) We applied SMOTE at 5000% so got almost 50% of each outcome, which would improve the quality of the prediction of the algorithms. Thus, to overcome the accuracy paradox we decided to use SMOTE. (Accuracy Paradox, n.d.)

Different techniques that work at different levels can be used for imbalanced data set classification (Patil and Sonavane, 2017). As given in Lopez et al. (2013; 2014) these levels include data level, procedure level and cost-sensitive level. In the data level, the sizes of the data sets are updated. For managing imbalanced Big Data sets mostly techniques at procedure level are used. Finally, the combined data and procedure level constitutes the cost-sensitive level.

At the data level there are three types of techniques, namely undersampling, oversampling and hybrid sampling all of which have pros and cons. While oversampling may produce noisy data, undersampling might cause to lose some useful data. For applying both techniques the easiest way is using the random approach (Batista, Prati, Monard, 2004). According to Lopez et al. (2013), oversampling results are found to have extra advantages when compared to undersampling techniques. One of the basic and most used oversampling technique is called Synthetic Minority Oversampling Technique (SMOTE) algorithm (Chawla et al., 2002). In SMOTE, ‘KNN’ nearest neighbors (KNN) are selected randomly to satisfy the oversampling rate.

Along with Random Oversampling, Adaptive Synthetic and Borderline SMOTE (Han, Wang, Mao, 2005) there are also some newer algorithms that may offer better predictive results in some cases. One is Box Drawings (Goh and Rudin, 2014), and another is Isolation Forests (Liu, Ting, Zhou, 2012).

Although these (especially borderline SMOTE, which does not create synthetic examples for noise instances, but concentrates its effort near the borderline, which in turn helps the decision function to create better boundaries between classes) are worth mentioning, (Dattagupta, 2017) using regular SMOTE was much preferable for us since it was natively present in Azure ML and tested thoroughly by many studies. These algorithms may become a subject of a future review, of course.

3. PROJECT STATEMENT AND METHODOLOGY

In this section, we state the objective and scope of the project. We briefly introduce our data set and explain the aims of Koalay's business model. Then, we present the methodology such as descriptive data analysis and model deployment for ranking the sales leads.

Since October 2017, when koalay.com extended its online products, they aim to increase their insurance sales. They decided to use Machine Learning tools to improve their insurance sales by understanding the client decision-making process and the relevant features. They also aim to rank their current sales leads on the ones who seem to be the most probable to buy. The ranking is performed using the knowledge gained with ML applied to the full sales leads. Then, by focusing on the ones predicted as sold but not sold, they start analyzing from the ones that scored highest, close to one.

3.1. Problem statement and Dataset

Koalay.com has a significant amount of data accumulating due to online quotes generated by 7/24. The dataset contains all online quotes generated from October, 1 2017 to June, 30 2018. It initially had almost 200k rows. Data cleaning and especially the removal of missing data rows took some time, and the data used for the analysis become 156k rows. The only major missing data removal was made due to the TEMINAT-ARAC DEGERI (value of the vehicle) column. Since this column is critical, we delete those 44k rows, which had NULL's as vehicle values, without hesitation. Also, nineteen columns were ready for analysis.

3.1.1 Data Cleaning and other pre-analysis steps

The SQL query of Koalay was designed to provide a smooth output. The major missing data was on the TEMINAT (vehicle value) column, and 44k rows were removed.

A second column, called “Trafik Hasar Kaydı,” provides the level of damage of a particular vehicle, and it is a numeric value between 1 and 7, 7 being the best value and 1 being the worst. A new car starts at level 4, and in the following years, if it has no damages, this value increases up to 7. Conversely, as it has more damage, it may become gradually even 1, which is the worst level of damage. We converted these column’s values to categorical variables by adding “A” as a prefix, so our categorical variables are A1, A2, A3, A4, A5, A6, A7. A7 and A4 are the most frequent ones. 29k rows had NULL in this column. We found out that this feature was not critical among others. Although we did not remove rows with NULL “Trafik Hasar Kaydı” values, we assigned the second popular value A4 to them, which was also the default value for a vehicle with no history.

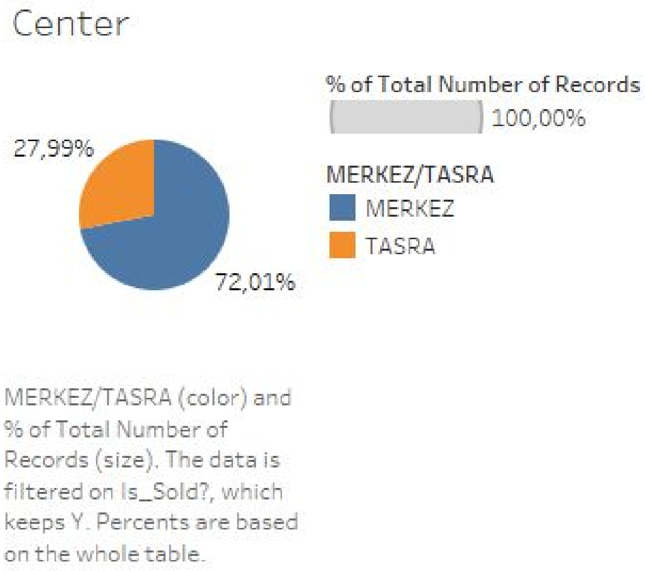
“Yas” column (age of the client) had 11k NULL values. We replaced them with the median age of 40.

Other smaller corrections were also made for AYD** - AYDIN, DEN**** - DENİZLİ, K.MARAŞ -KAHRAMANMARAŞ, İÇEL – MERSİN, IÇEL – MERSİN, MERSIN – MERSİN, DOĞUBEYAZIT – DOĞUBAYAZIT, EYÜP – EYÜPSULTAN. For these corrections, the latest version of the “City and District Database” of the Turkish General Directorate of Population and Citizenship Affairs was used. (İl, İlçe Kod Tablosu, NVI, 2016)

Age of client also had some flaws, few quotes that had very high client ages stated up to 136, were removed. They were not dependable, and it was clear that those are outliers or possibly errors.

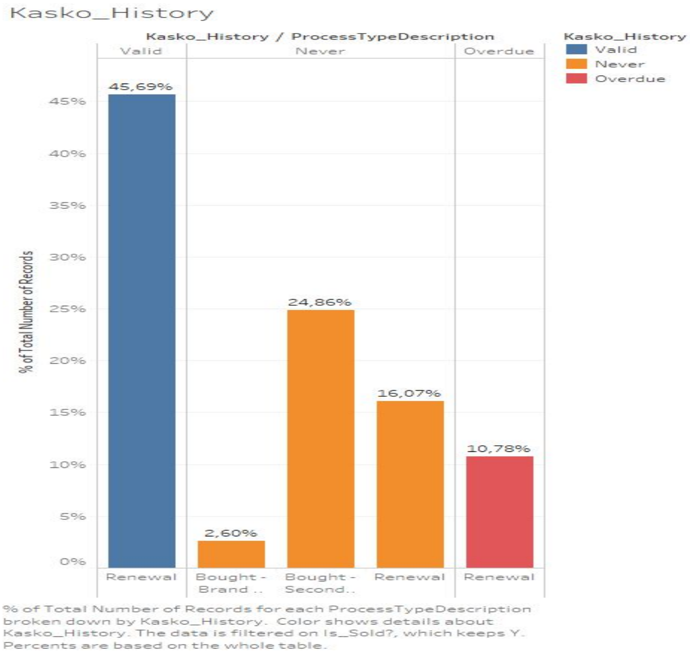
3.1.2 Visual analysis (EDA)

The dataset was imported into Tableau for visualization. After filtering the 3k sold quotes, we created the charts below. Thus, some facts were easily understood.



MERKEZ/TASRA (color) and % of Total Number of Records (size). The data is filtered on Is_Sold?, which keeps Y. Percents are based on the whole table.

Figure 1 – Location of policies sold

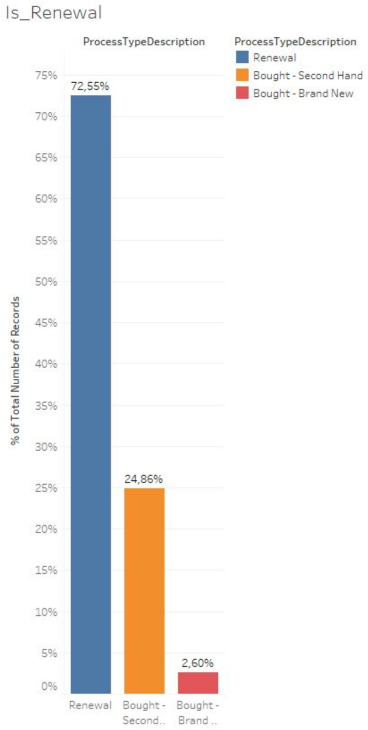


% of Total Number of Records for each ProcessTypeDescription broken down by Kasko_History. Color shows details about Kasko_History. The data is filtered on Is_Sold?, which keeps Y. Percents are based on the whole table.

Figure 2 – Kasko_History feature

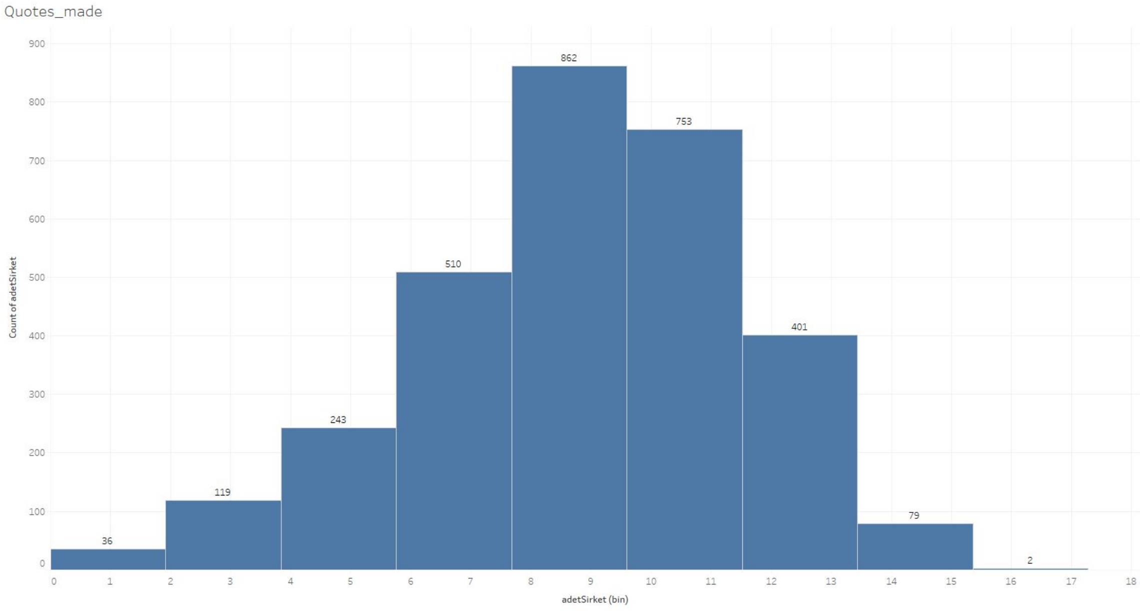
As seen from Figure 1, 72% of the policies sold were in City centers, in other words downtown (MERKEZ), and the 28% was in uptown (TASRA) addresses. As stated in Figure 2, 45% of the sold policies had a current valid policy, 44% of them never had a policy, and 11%

had an overdue policy. Of the vehicles that had no kasko history most were second-hand vehicles. Renewals followed, and the last group was brand new vehicles. This was in line with our expectations.



% of Total Number of Records for each ProcessTypeDescription. Color shows details about ProcessTypeDescription. The data is filtered on Is_Sold?, which keeps Y. Percents are based on the whole table.

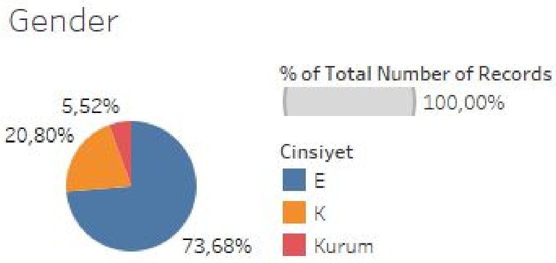
Figure 3 – ProcessTypeDescription feature



The trend of count of adetSirket for adetSirket (bin). The data is filtered on Is_Sold?, which keeps Y.

Figure 4 - The adetSirket feature shows the number of insurers in a quote

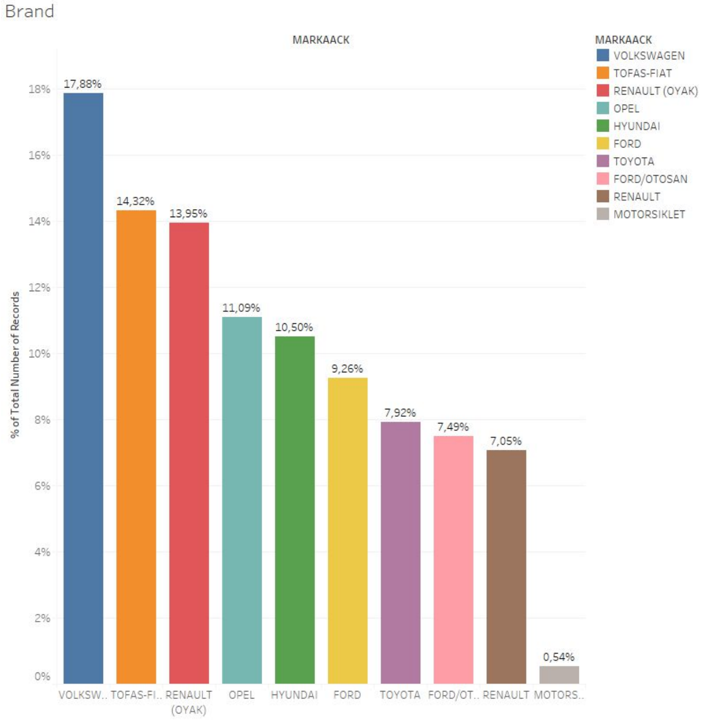
As can be seen on Figure 3, 72% of the sales were renewals, 25% were made for second-hand vehicles, and only 3% was made for brand new vehicles. Figure 4 states that the median number of insurers in a quote offered was nine, at a roughly standard deviation.



Cinsiyet (color) and % of Total Number of Records (size). The data is filtered on Is_Sold?, which keeps Y. Percents are based on the whole table.

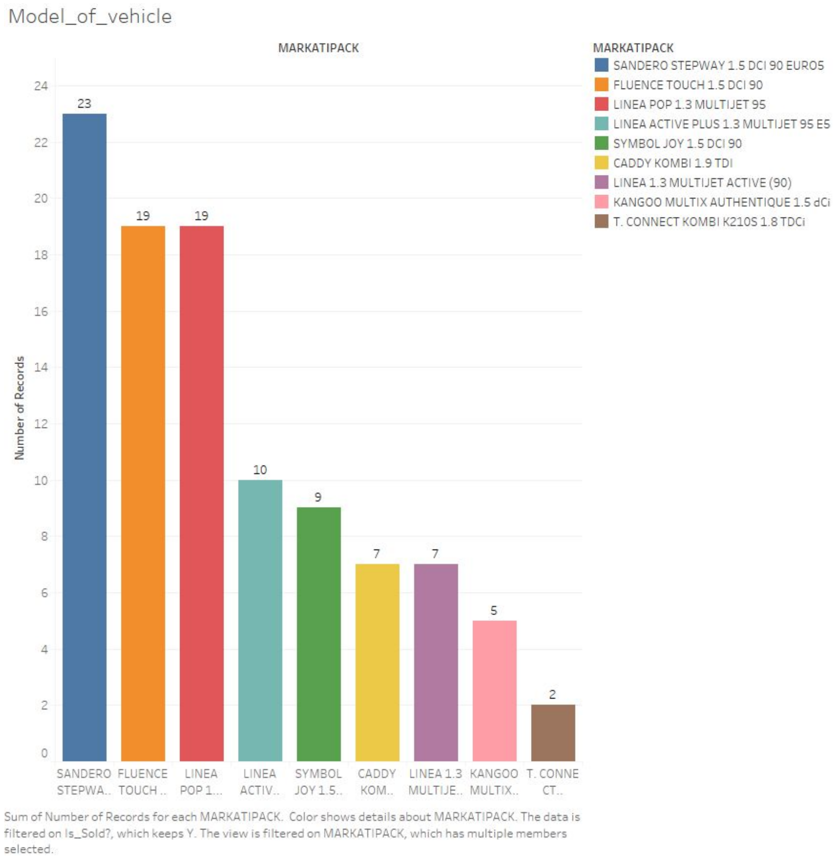
Figure 5 – Gender of the clients who bought policies

Most clients were males (E, 73%), Institutions (Kurum, 21%) were after women (K, 6%), as the last group of clients who bought policies.


Brand

% of Total Number of Records for each MARKAACK. Color shows details about MARKAACK. The data is filtered on Is_Sold?, which keeps Y. The view is filtered on MARKAACK, which has multiple members selected. Percents are based on the whole table.

Figure 6 – Most common brands of vehicles for which the policies were sold



Sum of Number of Records for each MARKATIPACK. Color shows details about MARKATIPACK. The data is filtered on Is_Sold?, which keeps Y. The view is filtered on MARKATIPACK, which has multiple members selected.

Figure 7 - Most common models of vehicles for which the policies were sold

Volkswagen was the most popular brand within the policies sold. Moreover, Dacia's Sandero Stepway 1.5 DCI was the most popular model within the policies sold.

We provide the feature descriptions of our data with sample values in Table 1.

Table 1. Feature descriptions and sample values

Feature	Type	Sample values
Tarih	Date, of the online quote generated	16.01.2018
EskiMusteri	String, existing client or not	Evet, Hayır
Cinsiyet	String, Gender of client or Institution	E, K, Kurum
Yas	Integer, Age of the client	44
IlAdi	String, the name of the city in which the vehicle is or will be registered	ADANA
IlceAdi	String, the name of the district in which the vehicle is or will be registered	YÜREĞİR
MERKEZ/TASRA	String, feature engineering. This feature was not a part of our database. Central or peripheric district.	MERKEZ(central), TASRA (peripheric)
ProcessTypeDescription	String; Renewal, Bought-New, Bought-Second Hand	Renewal
KModelYili	Integer, Model year of the vehicle	2008
MARKAACK	String, Brand (make) of the vehicle	HONDA
MARKATIPACK	String, Model of the vehicle	CIVIC 1.6 ELEGANCE
VehicleTypeName	String, Automobile-Pickups (90%) and others (10%)	Otomobil
TEMINAT	Integer, Value of the vehicle in TL	50779
Kasko_History	String; automobile insurance history of the vehicle (Never had, Overdue, Valid)	Overdue
KullanimTarzi	String, Usage Type of the vehicle; Private, Business, Official (public)	Özel (Private)
minPrim	Integer, Minimun premium quoted in TL	932
adetSirket	Integer, Number of quotes provided for that particular vehicle	5
Is_Sold?	String, Y or N (may convert to Boolean)	Y
TrafikHK	String, Traffic Damage Level (A1 lowest/no damage; A7 max damage history)	A7

4. EVALUATION OF THE INFORMATION GAIN UTILIZATION

After the analysis, the Permutation Feature Importance list was generated as below using the three models. The Two-Class Boosted Decision Tree, the Two-Class Neural Network and the Two-Class Locally Deep SVM.

Shortly we can state that if there is enough CPU power available, Two-Class Neural Network would be the best algorithm to stick with, if not the Two-Class Locally Deep SVM is the second best in performance, requiring much less CPU power and much less time with good enough prediction. The basic model of Two-Class Boosted Decision Tree provided less information gain.

As it can be seen in Table 2; the feature “EskiMusteri” was found highly important in all three models. This feature states if this client is an existing client or a new one. Existing clients have generated 80% of the sales made.

The “Kasko_History” feature, which states the automobile insurance history of the vehicle (Never had, Overdue, Valid), was also found highly important in all three models. As previously shown in Figure 3, 45% of sales were of Renewals of current policies, 44% was of vehicles that Never had an insurance policy. Moreover, remaining 11% was of vehicles with Overdue policies.

Neural Network and Locally Deep SVM models outperformed the other model on the “MARKATIPACK” feature, which states the model of the vehicle. This was one of our major concerns whether some makes and models of vehicles were more probable buyers of insurance policies. This will be the primary feature to follow in future evaluations, it may become more important.

Also, the features stating the registered city “IlAdi” and the district “IlceAdi” of the vehicle were found important by Neural Network and Locally Deep SVM. These features may be the second and third features to be scrutinized in future models along with the above mentioned “MARKATIPACK”.

Surprisingly “TEMINAT” feature, which is the value of the vehicle, had almost zero importance according to Neural Network and Locally Deep SVM models. The Decision Tree model found significant importance and we intuitively expected it might have significant importance. In the future, it will be tested again with these models using much more data.

Table 2. Feature importances of the three models used

Two - Class Decision Tree Feature Importances	Score	Two - Class Neural Network Feature Importances	Score	Two - Class Locally Deep SVM Feature Importances	Score
EskiMusteri	0,191241	MARKATIPACK	0,113706	EskiMusteri	0,195143
Kasko_History	0,052958	EskiMusteri	0,095514	IlAdi	0,132779
TEMINAT	0,052256	IlceAdi	0,093620	IlceAdi	0,095432
IlAdi	0,035975	IlAdi	0,076065	Kasko_History	0,082189
ProcessTypeDescription	0,030651	Kasko_History	0,062462	MERKEZ/TASRA	0,081552
minPrim	0,028822	Tarih	0,055734	ProcessTypeDescription	0,051276
Yas	0,025360	MERKEZ/TASRA	0,043585	MARKATIPACK	0,050639
IlceAdi	0,024185	ProcessTypeDescription	0,036775	adetSirket	0,030178
adetSirket	0,023025	adetSirket	0,033754	Tarih	0,018731
MARKATIPACK	0,021539	MARKAACK	0,023385	MARKAACK	0,006009
Cinsiyet	0,015677	KModelYili	0,011496	TrafikHK	0,00583
MERKEZ/TASRA	0,012052	TrafikHK	0,009341	VehicleTypeName	0,004034
KModelYili	0,010108	VehicleTypeName	0,008279	KModelYili	0,00338
Tarih	0,005291	Cinsiyet	0,005862	Cinsiyet	0,003217
TrafikHK	0,002613	Yas	0,001584	Yas	0,000833
MARKAACK	0,001323	KullanimTarzi	0,000033	KullanimTarzi	-
VehicleTypeName	0,000555	minPrim	-	minPrim	-
KullanimTarzi	-	TEMINAT	-0,000033	TEMINAT	-0,000049

After we completed the analysis using the three models, we obtained the confusion matrix data as below, thanks to SMOTE we got convincing precision scores of 89% or higher.

Again, the Two-Class Neural Network algorithm outperformed other models as seen below, with an AUC of up to 98% at 50% threshold.

Table 3. Confusion matrix of the test data from the three models used

Two- Class Boosted Decision Tree	Test Data 61k rows	Two- Class Neural Network	Test Data 61k rows	Two-Class Locally Deep SVM	Test Data 61k rows
TRUE Positive	FALSE Negative	TRUE Positive	FALSE Negative	TRUE Positive	FALSE Negative
27324	**3494**	**30386**	**432**	**28261**	**2557**
FALSE Positive	TRUE Negative	FALSE Positive	TRUE Negative	FALSE Positive	TRUE Negative
3372	**27047**	**953**	**29466**	**2934**	**27485**
Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
0,888	**0,89**	**0,977**	**0,97**	**0,91**	**0,906**
Recall	F1 Score	Recall	F1 Score	Recall	F1 Score
0,887	**0,888**	**0,986**	**0,978**	**0,917**	**0,911**
Threshold	AUC	Threshold	AUC	Threshold	AUC
0,5	**0,955**	**0,5**	**0,978**	**0,5**	**0,967**

Confusion matrix of the Two-Class Neural Network model is given in Table 4.

Table 4. Confusion matrix of the Neural Network model

Two- Class Neural Network	All Data 306k rows	T-C NN Training Dataset	Train Data 245k rows	Two- Class Neural Network	Test Data 61k rows
True Positive	False Negative	True Positive	False Negative	True Positive	False Negative
152821	**434**	**122435**	**2**	**30386**	**432**
False Positive	True Negative	False Positive	True Negative	False Positive	True Negative
954	**151978**	**1**	**122512**	**953**	**29466**
Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
0,995	**0,994**	**1.000**	**1.000**	**0,977**	**0,97**
Recall	F1 Score	Recall	F1 Score	Recall	F1 Score
0,997	**0,995**	**1.000**	**1.000**	**0,986**	**0,978**

4.1. Applying Scorings

In this section we try to apply the information gained to our full list of quotes given.

- We will start by filtering the quotes that were predicted as sold but not sold.
- In other words, we will deal with the highest scored False Positives for increasing sales.

4.2. Sales Lead Ranking Process

We, (koalay.com) as the owner of this data, are planning to publish and update the quarterly data on Azure and regularly utilize the latest rules for ranking the leads on an Azure web service. Then those leads will be provided to the sales staff for realizing the seemingly possible sales.

4.3. Actual Usage of Results

The result excel file shared with the sales team is also provided with our study, which was designed to provide benefit to the company by increasing sales. As it can easily be seen, these quotes were not sold but labeled as Y (can be sold) with scores up to 1 by our trained model. Some columns from a sample row is given in Table 5. The sales system can auto dial or e-mail the relevant client of the quote.

Table 5. Sample data from the result file

Tarih	EskiMusteri	Cinsiyet
29-06-18	Evet	K
Yas	**IlAdi**	**IlceAdi**
33	ZONGULDAK	MERKEZ
MERKEZ/TASRA	**ProcessTypeDescription**	**KModelYili**
TASRA	Renewal	2011
MARKAACK	**MARKATIPACK**	**VehicleTypeName**
TOFAS-FIAT	LINEA ACTIVE PLUS 1.3 MULTIJET (90)	Otomobil
TEMINAT	**Kasko_History**	**KullanimTarzi**
35636	Overdue	Özel
minPrim	**adetSirket**	**Is_Sold?**
420	11	N
TrafikHK	**Scored Labels**	**Scored Probabilities**
A7	Y	1

5. DISCUSSION & CONCLUSION

We spent a great amount of time for cleaning. Some features were obsolete in several thousand rows, so we removed them. Since our data is live and running every second, in future, it is possible to get a higher quality of input data thanks to our software development and database management efforts.

As we stated in the prior sections of our problem statement and aim, this study has the potential for further improvement via some factors like:

- Using more data, which is spread on at least consecutive twelve months,
- Improvising for adding more features by extra feature engineering,
- And an improvement in ML algorithms.
    - We are especially eagerly waiting for radical neural network designs to overcome many challenges and find out new patterns, relations.

Thus, we believe that this study has provided significant benefit for ranking our possible sales leads and we will continue to improve the pre and post processes as well. Please note that our analysis is limited with the seemingly vast but limited features, which may be considered as a possible weakness. If more features were available and we also generated more by feature engineering, it may lead to better prediction, so ranking performance.

We wished that we had some extra data on basic insurer files, such as the color of the vehicle. As Shin and Lee concluded in their 2013 dated study: “However, although we found out the fact that car color has something to do with car accident in the field of safety, more consideration of other properties should be researched further.” (Shin and Lee, 2013)

It may be possible that the color preference, may lead to some clue on the propensity for insurance sales. This sounds exciting especially for ranking quotes of new vehicles with no traffic and insurance history.

6. REFERENCES

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst. Appl., 78, 225-241. http://doi.org/10.1016/j.eswa.2017.02.017

Kang, S., Song J., (2018) Feature selection for continuous aggregate response and its application to auto insurance data. Expert Systems With Applications 93 (2018) 104–117 https://doi.org/10.1016/j.eswa.2017.10.007

McDonald S, Wren C. Multibrand pricing as a strategy for consumer search obfuscation in online markets. J Econ Manage Strat. 2018;27:171–187. https://doi.org/10.1111/jems.12239

Li, J., Fong, S., Wong, R.K., & Chu, V.W. (2018). Adaptive multi-objective swarm fusion for imbalanced data classification. Information Fusion, 39, 1-24. http://doi.org/10.1016/j.inffus.2017.03.007

Cui, Y., Tobossi, R., Vigouroux, O. (2018) Modelling customer online behaviours with neural networks: applications to conversion prediction and advertising retargeting. GMF Assurances, Groupe Covea https://arxiv.org/abs/1804.07669

Jurgenson, T. & Mansour, Y. (2018). Learning Decision Trees with Stochastic Linear Classifiers. Proceedings of Algorithmic Learning Theory, in PMLR 83:489-528 http://proceedings.mlr.press/v83/jurgenson18a.html

Lu, C., Ke, H., Zhang, G. et al. Memetic Comp. (2017). https://doi.org/10.1007/s12293-017-0236-3

Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. 2007. Learning on the border: active learning in imbalanced data classification. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM '07). ACM, New York, NY, USA, 127-136. https://doi.org/10.1145/1321440.1321461

Daniel Winkler, Markus Haltmeier, Manfred Kleidorfer, Wolfgang Rauch & Franz Tscheikner-Gratl (2018) Pipe failure modelling for water distribution networks using boosted decision trees, Structure and Infrastructure Engineering, 14:10, 1402-1411, https://doi.org/10.1080/15732479.2018.1443145

Volkovska, K. (2018) Modeling the Predictive Performance of Credit Scoring by Logistic Regression and Ensemble Learning. Master's Thesis, Tartu University, Estonia http://dspace.ut.ee/bitstream/handle/10062/61019/volkovska_kateryna.pdf

Rankings of Features (p119) Urszula Stańczyk, Lakhmi C. Jain (Eds.) Feature Selection for Data and Pattern Recognition Studies in Computational Intelligence vol. 584 Springer-Verlag, Germany, 2015 https://www.springer.com/gp/book/9783662456194

Telesure (2018) Retrieved from: https://www.telesure.com.tr/en/about-us
Koalay (2018) Retrieved from: https://telesure.com.tr/en/koalay

Lin, Weiwei & Wu, Ziming & Lin, Longxin & Wen, Angzhan & Li, Jin. (2017). An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2738069. http://dx.doi.org/10.1109/ACCESS.2017.2738069

Blagus, Lusa ; SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013; 14: 106. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3648438/

Fernandez, Garcia, Herrera, Chawla SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research 61 (Apr 2018) https://doi.org/10.1613/jair.1.11192

SMOTE (2018) Retrieved from: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/smote

Accuracy Paradox (n.d.) Retrieved from https://www.wikiwand.com/en/Accuracy_paradox

Patil and Sonavane Improved classification of large imbalanced data sets using rationalized technique J Big Data (2017) 4:49 https://doi.org/10.1186/s40537-017-0108-1

López V, et al. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. J Inform Sci. 2013;250:113–41. https://doi.org/10.1016/j.ins.2013.07.007

Rio S, Lopez V, Benitez J, Herrera F. On the use of MapReduce for imbalanced big data using Random Forest. J Inform Sci. 2014;285:112–37. http://dx.doi.org/10.1016/j.ins.2014.03.043

Batista G, Prati R, Monard M. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Expl Newslett. 2004;6:20–9. https://doi.org/10.1145/1007730.1007735

Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. J Artif. Intell. Res.2002;16:321–57. https://doi.org/10.1613/jair.953

Han H., Wang WY., Mao BH. (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang DS., Zhang XP., Huang GB. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg https://doi.org/10.1007/11538059_91

Box Drawings for Learning with Imbalanced Data." Siong Thye Goh and Cynthia Rudin. KDD-2014, August 24–27, 2014, New York, NY, USA https://arxiv.org/abs/1403.3378

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-Based Anomaly Detection. ACM Trans. Knowl. Discov. Data 6, 1, Article 3 (March 2012), 39 pages. http://dx.doi.org/10.1145/2133360.2133363

A Performance Comparison of Oversampling Methods for Data Generation in Imbalanced Learning Tasks, Dattagupta S. J., Nov. 2017 https://run.unl.pt/bitstream/10362/31307/1/TEGI0396.pdf

İl, İlçe Kod Tablosu, (2016) Retrieved from : https://www.nvi.gov.tr/hakkimizda/projeler/mernis/il-ilce-kod-tablosu

Shin, Seong-Yoon & Lee, Sangwon. (2013). Correlation between Car Accident and Car Color for Intelligent Service. Journal of Intelligence and Information Systems. http://dx.doi.org/10.13088/jiis.2013.19.4.011