Data analytics and machine learning for smart decision making in automotive sector

Hamid Ahaggach
LIB Laboratory, University of Burgundy, Dijon, France
Hamid.ahaggach@u-bourgogne.fr

Abstract. The objective of this thesis is to conduct scientific research on the use of data science and artificial intelligence techniques in the practices of automotive dealership companies to assist them in their decision-making processes and to use data-driven methods with modeling approaches for computing these enterprises. By proposing algorithms capable of continuously extracting relevant information from a diverse and multi-structured automotive environment. Due to the large amount of data available within these companies, we will develop algorithms to correctly assess the situation, suggest recommendations for decision-making, develop marketing strategies, and automate manual tasks that cost time, effort, and money.

Keywords: Data science, Artificial Intelligence, Decision-making, Marketing strategies.

1 Context and goals

In our research, we mainly focus on two key aspects of decision support for the automotive sector. First, we consider car commercialization. Here, the aim is to provide an aid to tackle the inventory problems of car dealers. Indeed, car dealerships buy cars from manufacturers and sell them to their customers to make a profit. Dealers cannot just send the unsold cars back to the manufacturer. Keeping unsold cars in the parking lots for a long time is extremely costly and may even threaten the financial prosperity of these companies. Therefore, they will have to find a way to get rid of these cars and get prepared to receive newer models. We propose to use data analytics (DA) and machine learning (ML) to predict the time required to sell car models. Characteristics of cars and sales history are taken into account to make ML models that are capable of making correct predictions in most cases.

The second aspect we consider is damage assessment. Car dealers import vehicles from the manufacturer's lot by ship, truck, or train. On arrival, all vehicles receive a detailed quality control to analyze damage incurred during transport and are stored in the car park. We aim to replace expert inspection, which costs time, effort, and money, with a fast and reliable automatic protocol.

The first work we have conducted in this track is the use of an ontology (OWL) to model the different damage caused to the car, according to the size and type of damage (Dents, scratches, paint damage, etc.) and the type of vehicle which the damage is to be evaluated. The ontology is built based on the knowledge of insurance experts and a large amount of data in the form of reports that they fill out. This work is the beginning of a project to assess the damage to the car and determine the price required to repair it, based on images captured by high-resolution cameras of vehicles from all angles using image-mining and ML tech-

niques, or based on the available textual data that describes the damage using Natural Language Processing (NLP) techniques and Named-entity recognition (NER).

2 Predicting car sale time

This section presents the problem, related work, suggested solutions, and preliminary results to assist car dealers in their efforts to solve vehicle inventory problems.

2.1 Related work and Problematic

Due to a lack of available public datasets, there is few research literature on vehicle sales prediction. Most existing works focus on sentiment analysis and the impact of economic factors on vehicle sales, without a comprehensive analysis of vehicle-related attributes. Pai, P. and Liu, C. [1] put forward a model for the prediction of vehicle sales by sentiment analysis of Twitter data and stock market values using least squares support vector regression.

Wijnhoven, F., and Plant, O. [2] test the predictive power of car sales by the ratio of positive to negative tweets, the total number of mentions, Google trends, and the percentage of negative comments. They report that social media sentiments have relatively very weak salience to improve predictions of car sales. Gao, J. et al. [3] proposed a hybrid optimization approach to forecast automobile sales in China by using gross domestic product, consumer price index, highway mileage, and automobile ownership. Wang, F. K., et al. [4] proposed a model based on monthly sales in Taiwan to select the most influential economic indicators such as oil price, current automobile sales, and exchange rate. However, these models only adopt economic indicators to predict nationwide sales. It is not enough to predict the car sales of different brands only based on regional economic indicators.

There are also many works on prediction models based on time-series data in various attributes such as commodity sales forecasting [6] financial market forecasting [7, 12, 13], weather and environmental state prediction [5, 14, 16]. But in our case time series cannot be used as a tool to predict car sales in the coming months, because the sales information, in our dataset, do not follow a precise pattern. For example, if we take a car *C* which rarely sells, where $\{c_1, c_2..., c_t\}$ are the monthly sale values, then most of the values are equal to zero. Given the sale values $\{c_s..., c_e\}$ over a period [s, e], where c_s and c_e are respectively the sale value of the start and the end of the period, if we train a network based on sequential models like (*LSTM*, *RNN* ...), or based on a statistical analysis model like *ARIMA* to predict Y_{t+1}= F ($c_s..., c_e$), the sale value for the next period $t + 1$, then we will get wrong results. Therefore, the research questions that can be asked are: (a) How can we forecast sales using the available data? (b) Which features are important to use in our model? (c) How can we use our model to develop better marketing strategies?

2.2 Methodology

The proposed solution is to use car characteristics to predict selling time. So we have a dataset of pairs D = $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $X = \{x_1, x_2 \ldots, x_n\}$ contains the characteristics of cars such as the color, the price, the power of the engine, also the entry date of the

vehicle, because we are using historical data and trying to make predictions to decrease the time to market, so time is a central variable in our model. While $Y= \{y_1, y_2..., y_n\}$ is the time taken to sell the vehicles which can be numerical or categorical. We need to find function $f(x_i) = y_i$ that will be able to predict the time needed to sell a car, this function can be any ML algorithm. The proposed methodology operates in three steps: (i) Data processing: This step is very important and sensitive because it directly affects the results. In our case, we noticed existing of missing data in different attributes, we treat the problem in different ways (Mode, Mean, Regression...) according to the type of attributes and their influence on car sale time via data correlation. In addition, some data elements are anomalous due to recording errors and must be filtered out. We also noticed the presence of data of the same type, but written in different formats data integration step is therefore required to address these kinds of problems. (ii) Dimensionality reduction: We need to select intrinsic features to achieve high prediction accuracy and reduce computation costs. We also need it to visualize data in a reduced space. *Forward selection* gives us a good result compared to other dimension reduction methods. This method keeps only the most important features in a dataset and eliminates the rest; In this case, the features are not transformed. While we used *Principal Component Analysis (PCA)* to visualize the vehicles in 2D space. (iii) Model training: We use "supervised" classification algorithms. We have compared several ML algorithms, but only the below 4 used in this article give a reasonable result. *(1) Support Vector Machines (SVM)* is the best-known form of kernel methods statistical theory of learning. This method searches for the hyperplane that separates samples of each class ensuring that the margin between the closest classes is maximal. *(2) Decision trees (DT)* depend on choosing which attributes to use first to build the tree. *(3) Random forests (RF)* operate by constructing a multitude of decision trees. The output of the RF is the class selected by most trees. (*4) K-Nearest Neighbors (KNN)* stores all available data and classifies new data based on similarity to its neighbors.

2.3 Experimental Results

To test the 4 algorithms and compare their prediction results, we use a large-scale dataset provided by two car dealership companies covering their car sales activities for the period between Oct. 2013 and Nov. 2021. The dataset has 33 attributes and more than 73200 entries. The dataset of the first company contains 40700 among these cars there are 18800 new cars and 21900 used cars, and for the second company, there are in total of 32500 cars, among which there are 18700 new cars and 14000 used cars. We have labeled our data as follows: 0, if selling the vehicle takes less than 3 months. 1, if selling the vehicle takes between 3 and 6 months. 2, if selling the vehicle takes between 6 and 8 months. 3, if selling the vehicle takes between 8 and 12 months. 4, if selling the vehicle takes more than 12 months. Our goal is to predict the time margin that a car will stay in stock before being sold, we will build two models for each company, one model for used cars and the other for new cars. The dataset is randomly split into two parts: the training set (80% of the dataset), is used to train and the test set (20% of the dataset) to evaluate our model using 10-Folds cross-validation. The accuracy and training time are considered as comparison criteria between algorithms on the test set. The accuracy in our case is defined as follows:

Accuracy = The number of well-classified cars / The total number of cars (1)

Table 1. Results of the prediction on the datasets of the two companies.

Company	Vehicle type	Metrics	Models			
			KNN	*SVM*	*DT*	*RF*
1	**VN**	Accuracy	0.971	0.951	**0.990**	**0.990**
		Training time	**0.096 secs**	8.431 secs	0.996 secs	0.140 secs
	VO	Accuracy	0.849	0.854	0.814	**0.863**
		Training time	0.058 secs	13.52 secs	**0.057 secs**	0.647 secs
2	**VN**	Accuracy	0.967	0.944	0.987389	**0.994**
		Training time	**0.079 secs**	20.76 secs	0.140 secs	1.145 secs
	VO	Accuracy	0.845	0.862	0.802	**0.870**
		Training time	**0.066 secs**	47.15 secs	0.187 secs	1.484 secs

Table 1 shows that the RF gives much better results in comparison with other models, and this is because RF is composed of several DT that collaborate. In the case of the first company, both the DT and RF give the same accuracy score on a new vehicle because the data in this case are easy to be discriminated by the decision tree. We also note that KNN generally gives good results because it is based on data. SVM takes a lot of time to learn in comparison with other models because generally maximization problems take more time and depend on the performance of the machine used to train the model.

2.4 Discussion and Perspectives

In this paper, we proposed to implement SVM, DT, KNN, and RF to predict the time required for dealers to sell cars. A large-scale car sales dataset provided by two multi-maker dealership companies has been pre-processed to complete missing data and identify the car characteristics that have the greatest impact on car sales. These predictions give companies better ideas about the commercialization of vehicles by providing the characteristics of the car to be purchased and the model will answer the time needed to sell it. This hence helps them to put the right marketing strategy to avoid buying cars that are not easy to sell. In future work, we intend to extend this work by using deep learning techniques because of the large amount of data and the number of characteristics available that could allow the application of such techniques and obtain good results. We also intend to use customer behavior analyses to build a recommendation system based on association rules, to target customers who can buy specific cars based on the profile of former customers.

3 Damage assessment

Damage assessment in general and damage assessment for cars, in particular, are difficult tasks because there are no specific criteria to assess the damage. We aim to model car damage using an ontology, due to the lack of previous work in the field of cars. Our ontology is based on the insurance experts' knowledge and their description reports. We model the description of all damages through define all concepts, data properties, and object

properties and also we take into account the type of car, type and severity of the damage, as well as the part of the car damaged. So that it is evaluated with the same price despite the different experts. This work is the beginning of a project to damage assessment using 9500 images of cars captured by high-resolution cameras through image mining techniques and ML. First, We will use semantic segmentation to identify automotive parts then we will recognize the damaged car parts and model the damage using the proposed ontology and finally estimate the price according to the price database. In the case of using text data of 23000 damage cases. We will use NLP techniques and NER based on rules or based on learning to extract the entities and the relationship between these entities and extract the useful information in the form of ontology to estimate the price.

3.1 Related work

Works that attempt to assess damage in the automotive field typically do not use ontologies, but rather employ a simple damage assessment (small, medium, major). For this reason, we cannot estimate the exact price necessary to repair the damage. They also do not take into account the location of the damage and the type of vehicle to be assessed, among these works we mention Waqas, U. et al. [8] who have proposed an image-based method of processing automobile insurance. In this regard, they consider the classification of the vehicle damage problem, where the classes include average damage, huge damage, and no damage, based on deep learning techniques, a MobileNet model with transfer learning for classification is proposed. Kyu, P. M., and Woraratpanya, K. [9] implement deep learning algorithms, VGG16 and VGG19, to detect and evaluate vehicle damage based on real-world datasets. Algorithms detect the damaged part of the vehicle and evaluate it by location and then its severity. Initially, CNN models are trained on the ImageNet dataset, followed by fine-tuning, because some of the classes can be very precise to accomplish tasks specified. The learning is then transferred into pre-trained VGG models and some techniques are used to improve the accuracy of the system. However, the severity is only evaluated on three levels (minor, moderate, and severe). Bandi, H. et al. [10] also perform the same work of assessing the extent of car damage using accurate convolutional neural networks, through a high-quality dataset that includes pivotal parameters such as location information and repair costs, but the assessment remains weak because the damage is not accurately described. While Singh, R. et al. [11] propose a comprehensive system to automate the damage assessment process. This system takes images of the damaged vehicle as input and gives relevant information such as damaged parts and provides an estimate of the extent of damage (no damage, light or severe) for each part. This serves as an indication of the then-estimated repair cost that will be used in deciding the amount of an insurance claim. Popular instance segmentation models such as Mask R-CNN, PANet, and a combination of these two have been used along with transfer learning based on the VGG16 network to perform various tasks of identifying and detecting different classes of fragments and damages.

3.2 Expected Result and Perspectives

This research is a step towards replacing the costly manual damage assessment with the automatic assessment of the damage. The proposed OWL describes accurately the damage

on the vehicle that consider the first step to estimate the cost of repair by analyzing the damaged car images and by analyzing the textual reports.

References

1. Pai, P. F., & Liu, C. H. (2018). Predicting vehicle sales by sentiment analysis of Twitter data and stock market values. IEEE Access, 6, 57655-57662.
2. Wijnhoven, F., & Plant, O. (2017). Sentiment analysis and Google trends data for predicting car sales.
3. Gao, J., Xie, Y., Gu, F., Xiao, W., Hu, J., & Yu, W. (2017). A hybrid optimization approach to forecast automobile sales of China. Advances in Mechanical Engineering, 9(8), 1687814017719422.
4. Wang, F. K., Chang, K. K., & Tzeng, C. W. (2011). Using adaptive network-based fuzzy inference system to forecast automobile sales. Expert Systems with Applications, 38(8), 10587-10593.
5. Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:1704.02971.
6. Zhao, K., & Wang, C. (2017). Sales forecast in E-commerce using convolutional neural network. arXiv preprint arXiv:1708.07946.
7. Elsworth, S., & Güttel, S. (2020). Time series forecasting using LSTM networks: A symbolic approach. arXiv preprint arXiv:2003.05672.
8. Waqas, U., Akram, N., Kim, S., Lee, D., & Jeon, J. (2020). Vehicle damage classification and fraudulent image detection including moiré effect using deep learning. In 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE) (pp. 1-5). IEEE.
9. Kyu, P. M., & Woraratpanya, K. (2020, July). Car damage detection and classification. In Proceedings of the 11th international conference on advances in information technology (pp. 1-6).
10. Bandi, H., Joshi, S., Bhagat, S., & Deshpande, A. (2021, June). Assessing car damage with convolutional neural networks. In 2021 International Conference on Communication information and Computing Technology (ICCICT) (pp. 1-5). IEEE.
11. Singh, R., Ayyar, M. P., Pavan, T. V. S., Gosain, S., & Shah, R. R. (2019, September). Automating car insurance claims using deep learning techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.
12. Hsu, M. W., Lessmann, S., Sung, M. C., Ma, T., & Johnson, J. E. (2016). Bridging the divide in financial market forecasting: machine learners vs. financial economists. Expert Systems with Applications, 61, 215-234.
13. Wang, J., Wang, J., Fang, W., & Niu, H. (2016). Financial time series prediction using elman recurrent random neural networks. Computational intelligence and neuroscience, 2016.
14. Sapankevych, N. I., & Sankar, R. (2009). Time series prediction using support vector machines: a survey. IEEE computational intelligence magazine, 4(2), 24-38.
15. Zheng, H., Yuan, J., & Chen, L. (2017). Short-term load forecasting using EMD-LSTM neural networks with a Xgboost algorithm for feature importance evaluation. Energies, 10(8), 1168.
16. Koochakpour, K., & Tarokh, M. J. (2016). Sales budget forecasting and revision by adaptive network fuzzy base inference system and optimization methods. Journal of Computer & Robotics, 9(1), 25-38.