ETL on Car Insurance's Price Data in Mexico City

1st Rodrigo Cadena Rodríguez
School of Engineering and Science
Tecnológico de Monterrey
Mexico City, Mexico
a1652141@tec.mx

2st Dr. Laura Hervert Escobar
School of Engineering and Science
Tecnológico de Monterrey
Mexico City, Mexico
laura.hervert@tec.mx

3st Neil Hernández Gress
School of Engineering and Science
Tecnológico de Monterrey
Mexico City, Mexico
ngress@tec.mx

***Abstract*—A car accident entails a series of damages that can be material up to the loss of life. According to the National Public Health Institute, Mexico is in seventh-place worldwide and third place in Latin America in the ranking of deaths caused by car accidents. Although the loss of life is irreparable, a car insurance allows to compensate the losses caused in this type of accident. In Mexico only 30% of people who own a car hires a car insurance, even when it is mandatory. The main reason is the perception of high prices compared to the benefits obtained. Currently, car insurance companies consider the same risk for all customers, while companies in other countries consider relevant variables to personalize risk using some machine learning models that have been proved to be efficient and permitted a low-cost premium based on users profile. In this way, this research represent an important part of a Thesis project for the *profiling of car insurance policy holders in Mexico using Data Science*. In this work, data of insurance policies prices for the most used cars in Mexico were obtained from 11 car insurance companies and processed for its further use with car accident data in Mexico in order to profile drivers and get a better pricing model for insurance companies.**

***Index Terms*—ETL, data extraction, data transforming, car insurance, auto insurance, data analytics**

I. Introduction

Car accidents in Mexico are the cause of an average of 24,000 deaths per year [5]. Although the loss of life is irreparable, a car insurance allows to compensate the losses caused in this type of accident.

Unfortunately, only a few drivers hire car insurance because of the price to pay as the main reason [6]. Despite that, most of the insurance companies in this country are not making an effort to enhance the methods used to calculate the price of their policies [7], opening a possibility for lower, fair and personalized prices. Companies in other countries are using models that have been proved to be efficient and permitted to revolutionize their business models [8]. This is why an effort has to be made in Mexico to transform the car insurance industry to better predict accidents for a given customer to offer a better price for most car owners.

Data Science is the one that will help us to give a solution to this problem. But what is Data Science? It has been defined in many ways over time. Jeff C. Wu used the term data science for the first time in 1997 as a modern name for statistics [1]. Joel Grus defines a data scientist as someone who extracts insights from messy data [2]. Provost and Fawcett defined data science at a high level as a set of fundamental principles that support and guide the principled extraction of information and knowledge from data [3]. Yan and Davis view data science as a discipline that provides theory, methodology, principles, and guidelines for the analysis of data for tools, values, or insights [1]. Drew Conway, defined it as the intersection of Hacking or Programming skills, Math & Statistics knowledge and Substantive expertise, where machine learning is in the intersection of programming skills and Math & Statistics knowledge. The list could continue growing, but all the definitions for data science can be simplified as the use of other disciplines, expertise and tools for extracting knowledge from data.

The extraction of knowledge is what we are going to perform over data gathered from Mexico's government site of open data. But it is not that simple as obtain data and perform a Machine Learning model to have a solution. In order to model the data, we should manipulate our data before. Imagine that you are going to cook lasagna, after you get all the ingredients, you don't just put them in the oven and the lasagna makes itself from nowhere. Before you put something in the oven, you need to prepare all the ingredients. Cook some meat, boil the pasta, chop tomatoes and so on. After you have all the ingredients ready in a container, now is the time to turn on the oven and put it inside. This analogy can be seen when we are working with data (the lasagna) and applying machine learning models to it (putting it inside the oven). After we get some data from an API, a Data Base or a CSV file on the internet, we have to prepare it in order to apply some ML models. This preparation could have many ways to be done (as for a lasagna recipe). But the main goal is to get the most insightful features so that the ones that are useless (eg. same value in all registers, few unbalanced values, too many null values, etc.)

The rest of this document is organized as follows: The methodology followed and data description is described in section II, then our results will be presented in section III and finally, in section IV we give our conclusions and future work is mentioned.

II. Method and Data

A. Methodology

As part of the entire project, we decided to use the Cross-Industry Process for Data Mining (CRISP-DM) standard process model. CRISP-DM has been proven to be useful for planning, documentation and communication of Data Mining projects [4]. The life cycle of this model consists of 6 steps that can be repeated over as many iterations as needed. Figure 1 shows the diagram of this cycle, where the order of the phases is as follows: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment.

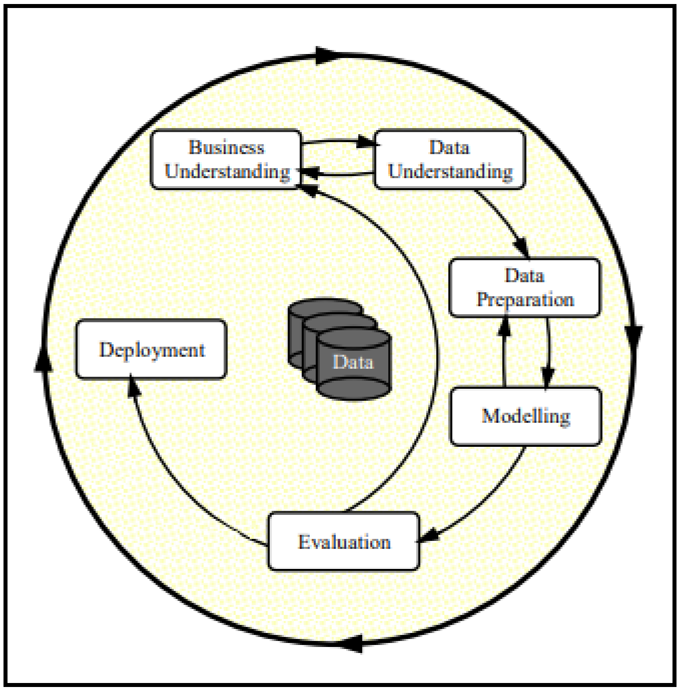

Fig. 1. CRISP-DM life cycle diagram [4]

1) Business Understanding: The main goal of this part is to understand the objectives and requirements that the project contains, so that a data mining problem can be defined in order to start planning the steps for giving a solution [4]. The intention of our project is to find a way to profile the risk of drivers in Mexico based on their sociodemographic data. After having a profile for a given group, the intention is to price that risk which could help insurance companies to lower their premium prices and as a consequence, increase the amount of insured drivers.

2) Data Understanding: This phase starts with the collection of data, which were built manually from the autocompara.com webpage, the resulting dataset was composed of 35,952 registers taking the 5 most used models in Mexico gathered from the INEGI's Sale to the public of light vehicles by brand, model, segment and country of origin dataset [9] and the 3 policy types: public liability, limited and wide coverage. For each model, 3 random versions were selected and the price from 11 insurance companies was registered for each coverage. Table I shows the description of the data in our built dataset.

3) Data Preparation: Here we performed a basic EDA, as the dataset was built by us, we knew that data were complete without empty values and almost balanced in all variables. Then, we encoded categorical data, so we ended up with 59 variables representing version, model, ages, years, and so on. Images 2 to 8 show how data are balanced in the entire population. We can see that the versions of cars are the only unbalanced variables, but this is because the models share versions when they belong to the same brand, eg. Nissan Versa Advance and Nissan March Advance. Also, there were some versions that were only present in some years, but not all of them.

TABLE I
TRAFFIC ACCIDENT DATASET ATTIBUTES (1)

No.	Attribute	Type	Possible Values
1	Brand	Qualitative - Nominal	Nissan, Volkswagen, GM
2	Model	Qualitative - Nominal	Versa, Aveo, March ...
3	Version	Qualitative - Nominal	Advance, Exclusive, Style, ...
4	Transmission	Qualitative - Nominal	Aut, Std
5	Year	Quantitative - Interval	2012-2017
6	Age	Quantitative - Ratio	18,25,30,...,50
7	Sex	Qualitative - Nominal	H, M
8	Policy	Qualitative - Nominal	Amplia, Limitada, RC
9	Insurance_company	Qualitative - Nominal	Mapfre, ABA, ...
10	Price	Quantitative - Ratio	3520.50, 4367.24, ...

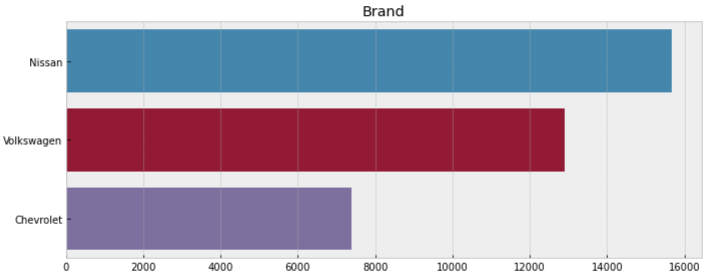

Fig. 2. Balance of brand values

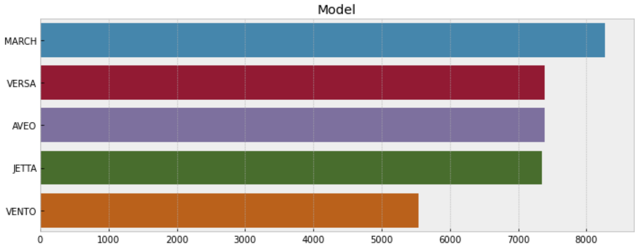

Fig. 3. Balance of model values

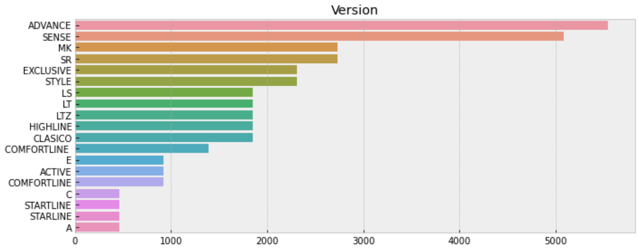

Fig. 4. Balance of version values

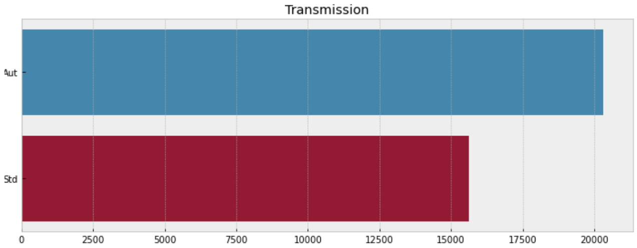

Fig. 5. Balance of transmission values

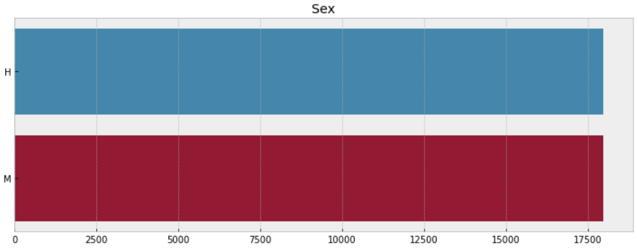

Fig. 6. Balance of sex values

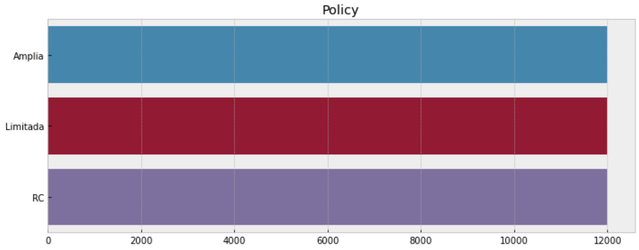

Fig. 7. Balance of policy values

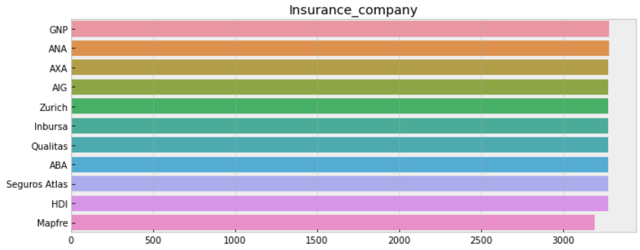

Fig. 8. Balance of insurance company values

4) Data Modelling: This section consists of the training of a model or models that will help develop a solution to the problem we want to solve. We used the data prepared from the last step and split into 80% training and 20% test, so the last one can be used in the evaluation phase. For this paper, we decided to perform a Decision Tree Regressor with a random state equal to 0. We used regression as the target variable (Price) was continuous.

5) Data Evaluation: This phase is focused on evaluating the performance of the models trained in the last section. We used cross validation with 10 layers to evaluate our Regression Decision Tree, and then we obtained the mean and standard deviation of the 10 scores obtained using this method.

III. Results

After performing a basic analysis over the Sex, Age and Price variables, we obtained Image 9, which is a plot of the people's age and age against the price of the policy. It can be seen that the higher the age of a person, the lesser the price to pay for an insurance policy, but the difference between male and female is almost insignificant.

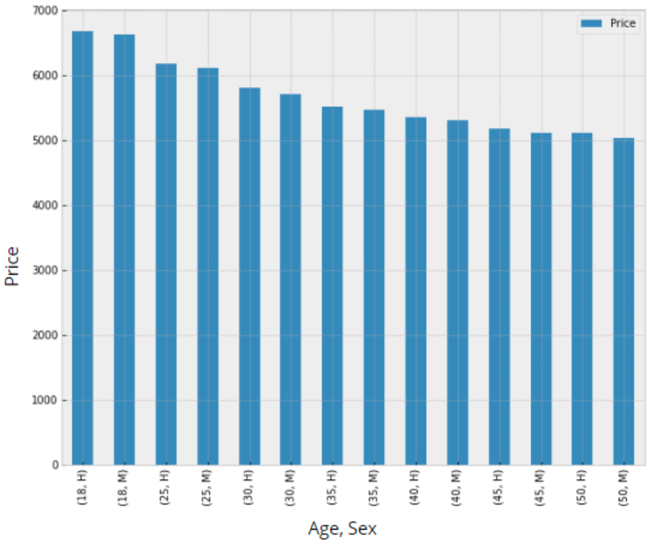

Fig. 9. Age and Sex vs Price

Regarding to the regression tree model applied to our data, we performed a cross validation with 10 layers over its results obtaining the following values: 0.87228591, 0.78412967, 0.7943663 , 0.8847869 , 0.81345142, 0.78035883, 0.84913777, 0.81357219, 0.86525554, 0.89928892. Then we got the mean of these 10 scores and it resulted in a value of 0.8356, which mean that our model had an accuracy of an 83%.

IV. Conclusion

Data Science has become I key part inside companies. It can help to improve several things inside them by improving their systems performance, the way they get new clients or maintain the existing ones by giving recommendations (streaming services) or just specific discounts based on their behaviour. In the case of our project, the intention is to give drivers a fair price, so people that drives carefully, will be able to pay less for an insurance policy. Sometimes it is better to construct your own dataset as it allows you to find some insights that cannot be seen in existing datasets. Here, we could see that sex is not a relevant feature as it does not cause a significant variation in the price of the policy. However, age is an important feature as it moves the price down when the person is older. A cause for this could be that young people tend to care less about their safety, drive faster and brake some traffic laws, whereas older people start to care about their families and beloved people, so they tend to drive better and safer within the law standards.

In future work, other Machine Learning models could be performed over our data to get better results in our predictions. Also, an effort must be made to find a way to match these data with the one that was gathered before about car accidents in the Mexican territory.

REFERENCES

[1] Yan, D.,Anddavis., G. E. A first course in data science.J. Stat. Educ.2019.

[2] Grus, J.Data Science from Scratch: First Principles with Python. O'Reilly, Beijing, 2015.

[3] PROVOST, F.,Andfawcett, T. Data science and its relationship to big data and data-drivendecision making.Big Data 1, 1 (2013), 51–59.

[4] Wirth, R. & Hipp, Jochen. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

[5] INSP, W. Meexico, septimo lugar mundial en siniestros viales. Tech. rep., Instituto Nacional DeSalud Publica, 2020.

[6] SALDIVAR, B.Costo, principal limitante del seguro para autos en Mexico: Swiss re. Available at https://0-search-proquest-com.biblioteca-ils.tec.mx/wire-feeds/costo-principal-limitante-del-seguro-para-autos/docview/2303623872/se-2?accountid=11643, Oct 09 2019.Copyright - El Economista, Mexico - Distributed by ContentEngine LLC; Ultima actualizacion - 2019-10-11.

[7] SANTANDER. Auto compara. Available at https://www.autocompara.com/.

[8] MEYER, J. D. The use of big data and artificial intelligence in insurance. Tech. rep., TheEuropean Consumer Organization, 2020.

[9] INEGI. Sale to the public of light vehicles by brand, model, segment and country of origin. Available at http://en.www.inegi.org.mx/contenidos/datosprimarios/iavl/tabulados/8_Ventas_serie.xlsx. INEGI, 2005-2022.