Diploma in Data Science, Artificial Intelligence and its Applications in Business and Economics

Final Project:
Real Estate Appraisal

Members: MANZANO, Carolina; MENGHI, Gonzalo.

TRABAJO FINAL: TASACIÓN DE INMUEBLES

Introduction

The real estate market is a key sector in the economy of any city or region. Both buyers and sellers require accurate and up-to-date information on property prices to make informed decisions. In this context, the ability to predict property prices can be highly valuable. This report will present a model for predicting property prices in Córdoba, a rapidly growing city that has become an increasingly attractive real estate market. The model is built using a database from July 2022 sourced from zonaprop.com.ar, where various individuals listed their properties for sale.

Description of the problem and objectives

Given the importance for buyers and sellers of real estate to know the publication price of their properties in order to be informed and increase the security in decision making, it is appropriate to develop a predictive model that allows the appraisal of real estate in the Province of Cordoba from the database obtained from the web within a time window.
It is in this sense that the development of the model that we present pursues some minimum desirable objectives:

- Our main motivation is to explore ways to build an accurate model for real estate price prediction, resulting in reliable estimates for buyers, sellers and stakeholders in the real estate market.
- We intend to identify the key characteristics of the properties that have the greatest impact on the listing price. Factors such as size of the property in m2, location, age and number of rooms such as bedrooms, bathrooms, garages, etc. are - to different extents and in different combinations - influential in the price.
- Adapt the model to different types of property descriptions so that its performance maintains most of its reliability over time.

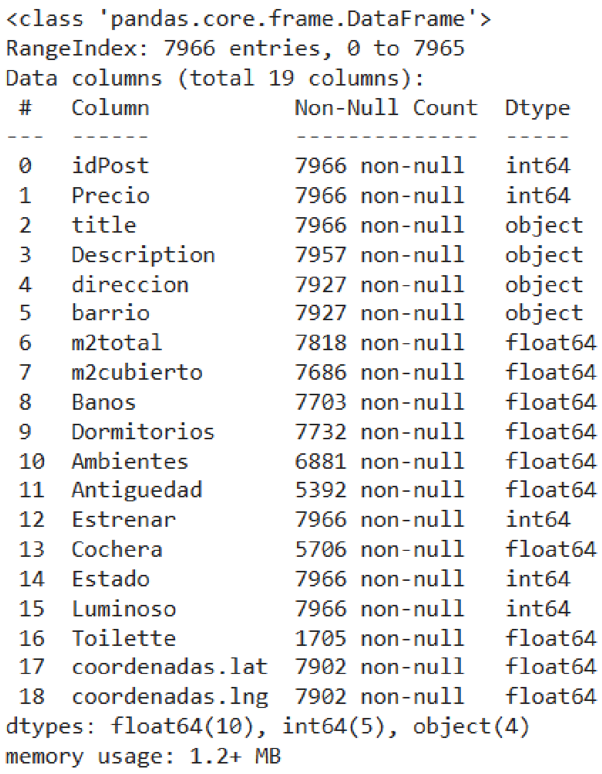
The development of the model maintains several business objectives, and it is from this that we infer that its implementation and publication will enable people:

- To be able to evaluate if the listing price of a property is fair and if it fits your needs and expectations in order to make more efficient decisions.
- Get more effective negotiations: Being informed will allow buyers to use this information to get a better price, while sellers will help them set a competitive price to maximize their profits.
- Being able to plan financially for both parties, as buyers can determine if the published price fits their budget and debt capacity, while sellers can estimate the return on their investment and plan for future transactions.

- To be able to identify investment opportunities and make strategic decisions based on price trends.

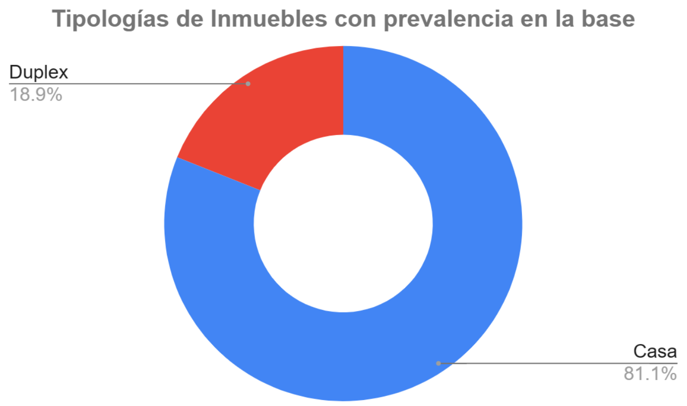
Exploratory data analysis and conclusions

La base de datos con la que trabajamos con datos de inmuebles publicados a la venta en julio 2022 en la Provincia de Córdoba cuenta con las características de la siguiente imágen:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7966 entries, 0 to 7965
Data columns (total 19 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   idPost            7966 non-null   int64  
 1   Precio            7966 non-null   int64  
 2   title             7966 non-null   object 
 3   Description       7957 non-null   object 
 4   direccion         7927 non-null   object 
 5   barrio            7927 non-null   object 
 6   m2total           7818 non-null   float64
 7   m2cubierto        7686 non-null   float64
 8   Banos             7703 non-null   float64
 9   Dormitorios       7732 non-null   float64
 10  Ambientes         6881 non-null   float64
 11  Antiguedad        5392 non-null   float64
 12  Estrenar          7966 non-null   int64  
 13  Cochera           5706 non-null   float64
 14  Estado            7966 non-null   int64  
 15  Luminoso          7966 non-null   int64  
 16  Toilette          1705 non-null   float64
 17  coordenadas.lat   7902 non-null   float64
 18  coordenadas.lng   7902 non-null   float64
dtypes: float64(10), int64(5), object(4)
memory usage: 1.2+ MB
```

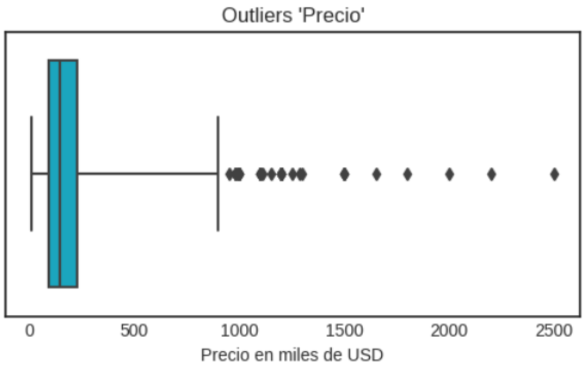
We started with the **Qualitative Variables Analysis** where we identified and eliminated 547 duplicate values of the variable "IdPost" and classified each publication by type of property published (house, duplex, apartment, local, land, etc.) based on the variable "title" where we decided to keep only those observations that correspond to the type "Casa" or "Duplex" and add them to a new column called "Type" since they cover a little more than 95% of the database.

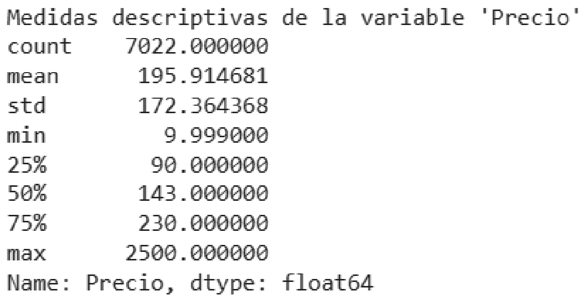


On the other hand, we characterized each publication according to the following words that we consider relevant due to the high frequency in which they are mentioned in the variable "Description" and we created a column for each one. These are: *“Suite”, “Pileta”, “Patio”, “Oportunidad”, “Reciclar”, “Country” y “Excelente ubicación”*.

When **analyzing the Quantitative Variables**, we faced the great challenge of making the data represent reasonable and logical values, since it is a database created by people who have different criteria and opinions at the time of completing the publication of their property for sale on the website, the information becomes abstract at times.

One of the cases, for example, was in the *variable "Precio"* where we had to omit the publications in which the sale value was less than USD 1,500 since we consider that it does not make any sense.

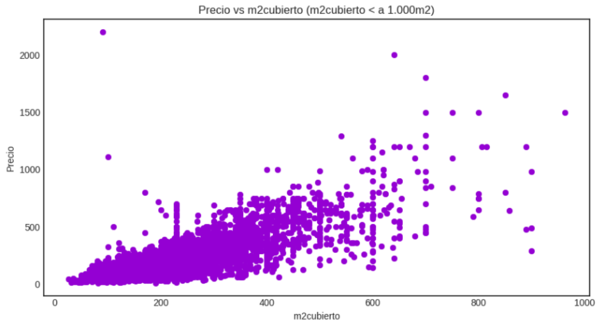


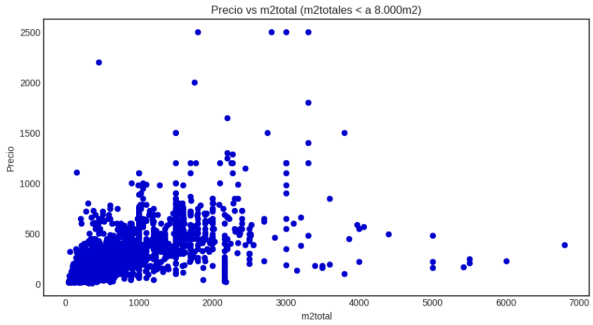


```
Medidas descriptivas de la variable 'Precio'
count    7022.000000
mean      195.914681
std       172.364368
min         9.999000
25%        90.000000
50%       143.000000
75%       230.000000
max      2500.000000
Name: Precio, dtype: float64
```

We discovered that in the *variable "m2total" and "m2cubierto"* there were publications with total m2 and covered m2 less than 50m2 (which we considered "outliers"), so we decided to replace the values not reported by the seller by the m2 reported in the variable "Description". Otherwise, by the average of total m2 and covered m2 respectively according to the type of property being sold.

In addition, after analyzing the behavior of the variable "*Precio*" with respect to "*m2totales*" and "*m2cubiertos*", we concluded that it would be a good decision to work with those publications in which the m2covered are less than 1,000 m2 and the m2total are less than 8,000 m2, since when exceeding these values we consider to be in the presence of atypical values for a property for sale within the typologies considered ("Casa" and "Dúplex").





With the rest of the variables we also worked with null values, as in the variable "*Baños*" where we assigned a value according to the information of the variable "*description*" and in case of not having the same, we assigned the value one (1) to it; and in "Dormitorios" where we also did it according to what was identified in the description in some cases and in the rest, we completed such data with the average of the variable. In addition, we decided to eliminate those publications that have more than six (6) bedrooms since we consider them "outliers" and have no relationship with the target variable.

In the variable "*Cochera*" and "*Antigüedad*", we replaced the missing values by the number zero (0) since it is assumed that if the seller of the property did not complete this information it is because the place does not have a garage or the property is brand new.

Finally, with respect to the variable "Toilette" and "Ambientes", after analyzing them and identifying the number of missing values and taking into account that they are fields that become abstract and can be freely interpreted by each person who makes a publication, we decided to disregard them from our analysis.

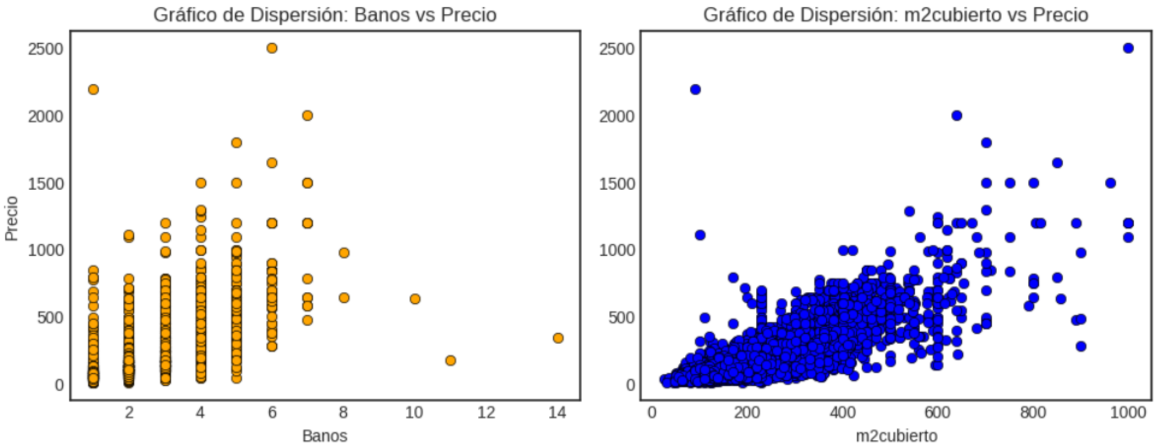
Based on these modifications, we created and saved a new database in order to perform the data modeling.

Model applied, Metric selected, Model chosen

After the exploratory analysis and data cleaning, we analyzed the correlation of the variables in the data set and observed that the most correlated variables with the target variable were the following:

- ***"m2cubierto"*** 80% positively
- ***"m2total"*** 64% positively
- ***"Banos"*** 63% positively

In addition, we observed a high correlation between the variable "Bathrooms" and "m2covered" where after analyzing it we conclude that it is normal and would not imply a multicollinearity problem; and in "m2totals" and "m2covered" where we could keep the variable "m2covered" and disregard the "m2totals" since keeping only the first one we would be doing a good analysis.



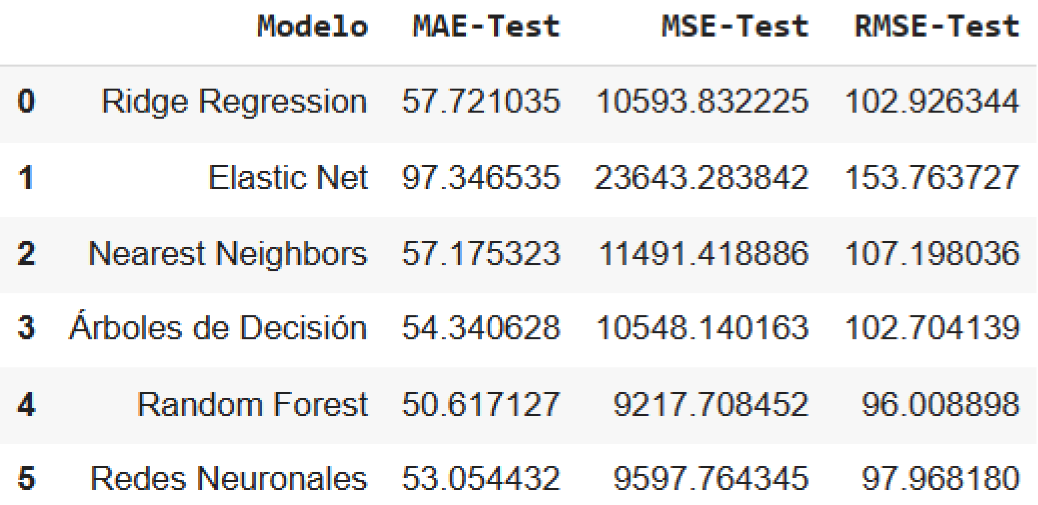
After this, we used Scikit-Learn's **StandardScaler()** function to scale the variables in our data set and tested the following models:

- Multiple Linear Regression Analysis
- Nearest Neighbors
- Decision trees for regression problems
- Random Forest
- Neural Networks

For each of the modeling techniques we follow the following steps to build each of the models:

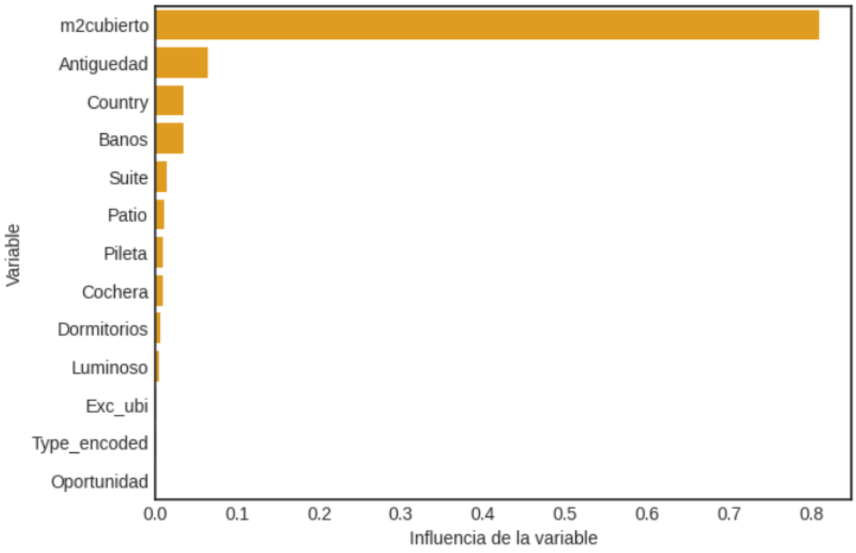
- Choose an algorithm that implements the corresponding technique.
- Find an effective parameter combination for the chosen algorithm
- Create a model using the parameters found
- Train the model on the training data set
- Test the model on the test dataset and obtain results

From the results obtained by each model, we present a table showing the evaluations of each model when applied to the X_test test set:



	Modelo	MAE-Test	MSE-Test	RMSE-Test
0	Ridge Regression	57.721035	10593.832225	102.926344
1	Elastic Net	97.346535	23643.283842	153.763727
2	Nearest Neighbors	57.175323	11491.418886	107.198036
3	Árboles de Decisión	54.340628	10548.140163	102.704139
4	Random Forest	50.617127	9217.708452	96.008898
5	Redes Neuronales	53.054432	9597.764345	97.968180

When selecting the optimal model, it is essential to rely on appropriate performance metrics so we decided to choose the **MAE (Mean Absolute Error)** metric since we are looking for robustness against outliers and we want to know the average deviation of the predictions.
Analyzing the results obtained, we observe that the "**Random Forest**" model is the one that presents the lowest MAE value at 50.6171 which means that, on average, the model will predict a higher or lower real value by 50.6171 thousand dollars (USD 50,617.10).
On the other hand, we observed which are the most important variables for this last model and concluded that definitely the most influential variable is "*m2cubierto*".



Final conclusions

There are several aspects of this research that deserve a mention and we intend to extract the most significant aspects of each stage of this project. Before starting the review we want to be categorical with this statement; we collected enough evidence and experience to feel satisfied with the work and the results presented here, especially if we consider that our objective was focused on the exploration of models. And beyond what exclusively concerns the research topic, we faced an everyday problem of the socio-economic life of any city in Argentina, something as essential as defining the monetary value of a good.

We started the challenge by taking data from a public website, it is very likely that by replicating the experience we can afford to work from the acquisition of underutilized data to add value from an exhaustive and careful process of cleaning and subsequent EDA. We learned to work on building an iterative model that explores as many possibilities as our creativity allows. Adding value does not go unnoticed, we took away with us the possibility of demonstrating how interesting the data that is in front of everyone's eyes can become.

From the beginning we noticed that the dataset did not skimp on problems in terms of human error and/or data that defy intuition. As it is not always easy to discriminate whether it is useful information or trash data, we understood the need for a statistical criterion to normalize the sample. We worked on cleaning by removing duplicate values by removing repeated IDs, outliers by analyzing the median of the variables and the jumps in the interquartile range, we also removed variables because they were affected by a large number of missing and abstract values (as was the variable "Toilette" and problems in finding criteria to count the number of rooms). Our difficulties in this aspect were proven at the time of the normality and heteroscedasticity test of our model. We applied Jarque-Bera tests to the residuals, D'Agostino's omnibus K2 test and an additional Kolmogorov-Smirnov test, all concluding that the distribution of the errors does not conform to a normal distribution, with p-values much lower than the standard, so we rejected the normality hypothesis.

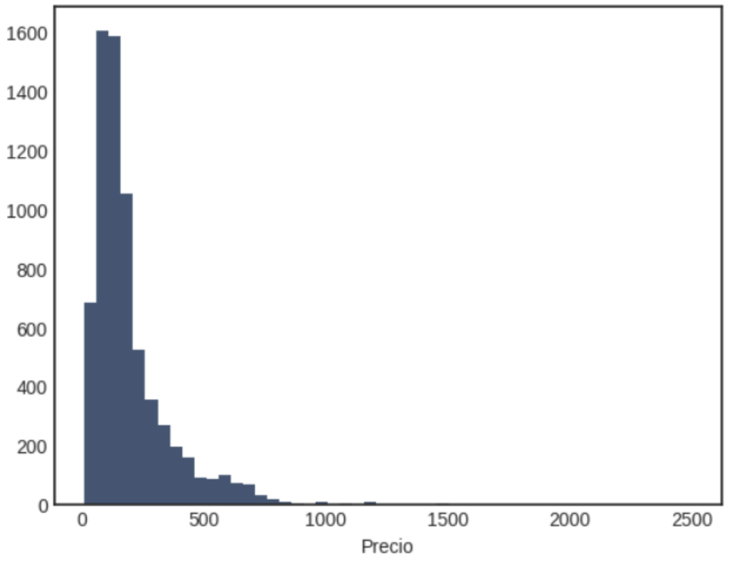
Confirming the importance of the variable "*m2cubiertos*" with data was a significant finding, from the correlation analysis of variables we quickly understood that it would become the most important of our project, later endorsed from the analysis of influence of variables. Thus, taking this variable together with: *"Banos", "Dormitorios", "Antiguedad", "Cochera", "Luminoso", "Suite", "Pileta", "Patio", "Oportunidad", "Country", "Exc_ubi", "Type"* we managed to explain just over 70% of the price variability.

It is a good time to think about our greatest opportunity for improvement: the geographical analysis of the published properties. It is plausible to think that in addition to the amount of m2cubiertos of a property, it also matters in which zones they are located. Although we started working on the creation of geographic zones, we preferred to spend time iterating from the information we had available.

That is why in the next stage we normalized the variables using Scikit-Learn's StandardScaler(), which allowed us to proceed with an objective comparison of different predictive models: from Multiple Linear Regression to more complex methods such as Neural Networks. For each one, we searched for the best parameter settings and evaluated their performance, establishing a robust framework for the training and testing phases. The results obtained bring us closer to an effective predictive model. We have built several predictive models, including *Random Forest, Neural Networks, Decision Trees, Nearest Neighbors, Ridge Multiple Regression and Elastic Net Multiple Regression*, and evaluated them in terms of performance and predictive ability.

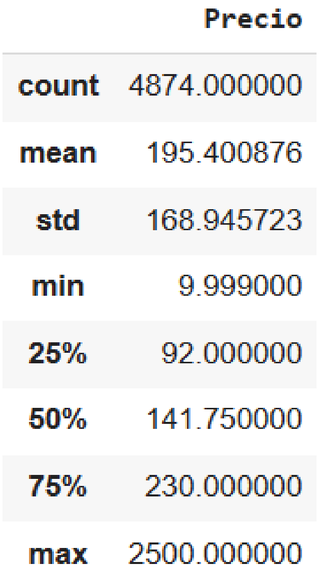
As a result of this analysis, we have identified that the **Random Forest Model** excels in terms of accuracy in predicting housing prices in Cordoba. But we also analyzed that the best models are those that work from nonlinear supervised learning methods. The variation of MAE among our tests suggests to us that it is the best at capturing complexities and nonlinear patterns in price formation and avoiding along the way some over-fitting problems.

The methodological decision that led us to consider that our model is reasonable was that, having a mean absolute error (MAE) of our selected model (Random Forest) of 50,617.1 USD, and considering that the mean sales price in our database is 195,400.88 USD, the median is 141,750.00 USD, and also the first quartile is 92,000 USD (indicating that 75% of the data exceeds this value), the error level is reasonable for being in tune with the characteristics of the local real estate market.



Elaboración propia en base a los datos proporcionados por Zonajob del año 2022.

Precio expresado en miles de USD



	Precio
count	4874.000000
mean	195.400876
std	168.945723
min	9.999000
25%	92.000000
50%	141.750000
75%	230.000000
max	2500.000000

We do not want to overlook other aspects that, like the incorporation of a geographic analysis criterion for the price variable and revising the rejection of the normality and heteroscedasticity hypothesis, would improve the performance of our model. We wonder

about the possibility of further experimenting, for example, by filtering the dataset from different ranges of sales prices -lower than the mean or higher than the mean- and interpreting the performance of the model in each scenario. Of course, another point to consider is that we work from a temporal snapshot of one month, an interesting way to go and strengthen the work is to extend the time window so that the model becomes much more flexible in discriminating seasons and general market movements.

We chose to leave for last another type of questions - not less important - that have to do with our context: how do (somewhat more) subjective factors such as economic instability, seller desperation or price exchange for cash payments influence the price of real estate, and what we find most exciting, how to capture them as variables for our model? We note some trends in proxy variables such as "opportunity" or "owner sells" to which we will not assign direct importance, but which allow us to identify (if we use them judiciously) other more meaningful relationships. We temporarily settle for a propositional answer: "*iterate and you will succeed*".