Real estate market in Barcelona

Architecture Decision Document

Oscar de Dios

June 10, 2020

Architectural Components overview

Data Source

Definition

For developing of any project the data is the most important element. High quality is mandatory and huge amount of data better than a few records to improve calculation process.

In order to get relevant information I checked the public data available from different sources. I visited a lot of sites but in more cases the free information were poor. Finally I worked with this data sources:

- Related to real estate market sales I collected the information from one of the most popular real estate sites on line in Spain (https://www.idealista.com). In this site I discovered a specific place with the information about pricing evolution over years (https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/cataluna/barcelona-provincia/barcelona/) I developed a little "web scrapping" code but unfortunately this site are protected of this type of code (Congrats site admin!). Finally I decided to install a Firefox plugin called Table2Clipboard that allows table capturing and export easily from HTML code.

Technology choice

- Anaconda Navigator, jupyter notebooks
- Firefox Plugin Table2Clipboard..
- Python, scikit-learn, Pandas, matplotlib, plotly, keras

Data integration

Technology choice

- **Anaconda Navigator**

  Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS, and Linux. https://docs.anaconda.com/anaconda/navigator/

- **Jupyter Notebooks**

  The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. https://jupyter-notebook.readthedocs.io/en/stable/notebook.html

- **Firefox Plugin Table2Cliplboard**

  Allow to copy to clipboard an HTML table rows/columns selection correctly formatted. https://addons.mozilla.org/es/firefox/addon/table2clipboard2/

- **Python**

  Python is a programming language that lets you work more quickly and integrate your systems more effectively. https://www.python.org/

- **scikit-learn** Simple and efficient tools for predictive data analysis. https://scikit-learn.org/stable/

- **Pandas**

  Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python.

  https://pandas.pydata.org/

- **Matplotlib**

  Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python https://matplotlib.org/

- **Plotly**

  plotly.py is an interactive, open-source, and browser-based graphing library for Python. plotly.py

- **Keras**

  Keras is an API designed for human beings, not machines. Keras follows best practices for reducing cognitive load: it offers consistent & simple APIs, it minimizes the number of user

actions required for common use cases, and it provides clear & actionable error messages. It also has extensive documentation and developer guides. https://keras.io/

Data repository

Technology choice for data storage are CSV files in local storage drive. Due to the type of information treatment we only need 11 different files with a few records on in. it is not needed a complex infrastructure storage. Further versions may require changes on data repository

Data discovery and exploration

There are a lot of options for visualizations but in this case we must follow some simple rules:

- What type of visualization are needed?
- The visualizations requires interactivity?
- Keep it simple

Technology choice

- **scikit-learn**

  Simple and efficient tools for predictive data analysis. https://scikit-learn.org/stable/

- **Pandas**

  Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python.

  https://pandas.pydata.org/

- **Matplotlib**

  Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python https://matplotlib.org/

- **Plotly**

  plotly.py is an interactive, open-source, and browser-based graphing library for Python. plotly.py

Actionable insights

- **Python**

  Python is a programming language cleaner than others and easy to work with. In this case I been using Jupyter Notebooks with Anaconda Navigator. The power of Python language going together with Pandas dataframe gives the consultant the perfect tool to be focused on needs.
  IBM Cloud via IBM Watson Studio is another powerful cloud tool that can be used with Jupyter Notebooks. I found some problems with the free version limitations, and decided to move to Anaconda Navigator on local infrastructure.

- **TensorFlow & Keras**

  Tensorflow is a great library to learn and dive into deep learning. For the project I selected keras sequential model to implement a deep learning algorithm for our application. Keras is easier to understand and build a deep learning model.

Model definition

I decided to use a recurrent Neural Network (RNN) based on Long Short-Term Memory Network (or LSTM network). It is a type of recurrent neural network to analyse sequence data. It learns input data by iterating the sequence elements and acquires state information regarding the checked part of the elements. Based on the learned data, it predicts the next item in the sequence. You can reach more information here: Time Series prediction LSTM

Model Evaluation

The model I developed was linked to time series. After reading a lot of papers I found the best summary on evaluating time series forecast on Rob Hyndman's site. I had much better success with the rooted mean square error (RMSE).

Model Deployment

For the model deployment Jupyter Notebook was the selected tool.