EXECUTIVE SUMMARY

The aim of this project is to develop an automatic valuation model for the prediction of real estate prices in three major cities in Spain including Barcelona, Madrid and Valencia in order to provide important information on the value of a property to the buyers, sellers, investors, and other property agents. Since it is crucial to obtain realistic estimations to determine the fair value, appraisals, and quick sales in real estate markets, this project used three primary datasets from Idealista, including 189,923 housing units in the cities with more than 40 variables, which were systematically cleaned and analysed. These datasets provided information on the characteristics of houses, points of interest, and boundaries of neighbourhoods. Additionally, secondary data was also incorporated to augment the baseline model at hand and are primarily related to the healthcare and education sectors.

According to PwC, while 91% of real estate investors recognize the potential of data analytics, only 39% effectively use them. The real estate market should therefore make use of the advancements in machine learning and artificial intelligence to drive investors' profits while remaining competitive. Hence, this gap constitutes a huge opportunity through which Idealista can penetrate the market of real estate valuation with an AI-based recommendation tool, "Idealista IQ", in order to enhance the efficiency of investment decisions by means of a recommendation engine. This one can be trained via a collaborative filtering algorithm and potentially increase the return on investment by as much as 15%. Idealista IQ offers users tailored investment insights: comparing historical prices, assessing the potential for rental income, and analysing market trends, as well as individual profile information and past behaviours. Besides, Idealista IQ by design should be much more user-friendly and would add a lot of value to the current Idealista Inversores tool and eventually offer detailed reports with a precise data-driven recommendation engine. It understands the correlations between preferences of individual user groups by using all user interactions with the website, including browsing history and ratings. Having taken into account more data, from the trends in the market to the statistics of a population, considering a neighbourhood, square metres, or a parking space, Idealista IQ will become the Netflix of the real estate market by offering a very strong recommendation system to investors who struggle to find profitable investment opportunities. On another note, generative AI will be leveraged to offer special domain expertise by providing a smart agent that would be readily available for help whenever a problem arises.

Understanding and exploring data is the first step in the preparation for any kind of predictive modelling work, and as a result in this case, it was clear that property pricing was correlated to such features as the area of construction, number of rooms, and distance from metro stations. Readiness of data and exploratory data analysis are the basic inputs of any predictive modelling exercise. Here, EDA showed important relations between property prices and features like constructed area, the number of rooms, and proximity to metro stations. Preliminary analysis showed that variables like price, square metres, a number of rooms, and bathrooms contain outliers, which were analysed and addressed. Handling duplicate ID values involved dropping the entries to have unique listings and, therefore, improved the

accuracy in the dataset. Additionally, missing values were also imputed by using the mean or the mode for categorical variables to ensure the robustness and reliability of the datasets. Outlier treatment involved limiting the ranges of some variables, ranging from values that would be detrimental to the model's prediction accuracy. This detail in data preparation ensured that the predictive models would be true and reliable. The distribution of the average unit-price of the properties across different neighbourhoods in each city also reflected the variety in the range of the prices based on location.

Various algorithms for machine learning were tested, including linear regression, random forest, XGBoost, and LightGBM. All models were placed on different data transformations to get the best results in addition to feature engineering. Among them, the most accurate model obtained was due to using XGBoost with Box-Cox transformation, which provided a MAPE of 0.87% for Barcelona, 0.24% for Madrid, and 0.60% for Valencia.

The accuracy obtained by this model was quite acceptable, considering our metric of interest for our case: the MAPE. However, some of the constraints that Idealista IQ will probably face are managing data privacy for the sake of compliance with protection laws, robust ETL processes in the mantle of data quality, and decreasing biases within a recommendation engine itself. Additionally, it must scale an infrastructure capable of handling huge volumes of data and user bases with continued monitoring and updates as the market does change. Furthermore, it shall show users the risks associated with realty investing in order to avoid developing dependency on its recommendations. In the future, Idealista IQ should partner up with financial institutions to provide seamless financing options and make use of blockchain for high transactional security. It should also enable AR and VR features that permit virtual property tours. To this, predictive analytics on market trends and a focus on sustainability in the form of highlighting eco-friendly properties should be added. These enhancements will further consolidate Idealista's position as an industry leader in Spain, providing full-feature and more intuitive tools with which to enhance investment decisions and improve client satisfaction.

In order to conclude, we can affirm that this process would require careful consideration and execution to leverage such analytical technology while ensuring conformance to regulatory standards. The incorporation of Idealista IQ with financial institutions and clients might be able to facilitate the overall financing process, hence making real estate investments more convenient for those interested. As briefly mentioned above, the tool will not only be able to create strong models of evaluation but also provides an all-inclusive platform that caters to the different requirements of its users. This aspect could be considered the key factor in differentiating such insights tool from others, creating a competitive advantage for success over the long term in the real estate market in the area of property valuation and insights investments.