Selling Smarter: Using Machine Learning to Anticipate Real Estate Pricing Outcomes

BSIM7217 assignment

Context

The Australian real estate market is one of the most dynamic and competitive in the world, offering a wide range of properties to both buyers and sellers. For homeowners looking to sell, setting the right price is a critical—and often emotional—decision. After all, property transactions are among the most significant financial events in a person's life.

Sellers typically set a listing price based on what they believe their home is worth and what the market might bear. But things don’t always go as planned. Some properties attract intense buyer interest and sell for more than the asking price. Others fall short, forcing the seller to accept less than they’d hoped.

If sellers had a way to estimate in advance whether their listed price is likely to be exceeded or undercut, they could make more informed pricing decisions, better manage expectations, and potentially maximize their return.

In this assignment, your task is to build a **binary** classification model that predicts whether a property will be sold at a higher or lower price than the advertised price set by the seller.

Target Variable

The target variable price_outcome indicates whether a property was sold at a higher, equal or lower price compared to the listing price.

The values in the price_outcome column are:

- Higher: Sold price is greater than the listed price
- Equal: Sold price is the same as the listed price
- Lower: Sold price is equal to or less than the listed price

This is a **binary** classification problem; therefore, you should not include any data where the target value is ‘Equal’. Your model should learn to predict this outcome using the available features of each property outlined below.

Dataset

You are provided with a dataset of 6,957 recently sold properties, between February 2022 and February 2023. The predictor variables are:

1. property_address: the address of the property
2. property_suburb : The suburb the property resides in
3. property_state : The state which the property resides in
4. listing_description: The description of the house provided on the listing

5. listed_date: The date the property was listed for sale
6. listed_price: The price the property was listed for
7. days_on_market: The number of days the property was on the market
8. number_of_beds: The number of bedrooms on the property
9. number_of_baths: The number of bathrooms on the property
10. number_of_parks: The number of parking spots on the property
11. property_size: The size of the property in square meters
12. property_classification: The type of property (House/Unit/Land)
13. property_sub_classification: The sub-type of the property
14. suburb_days_on_market: The average days in market that a property is on sale for in a suburb
15. suburb_median_price: The average median property price in a suburb

Deliverables

You must submit the following:

- ✓ A Jupyter Notebook (via the Assignment Submission link).

The aim of this assignment is to provide you with experience in the steps involved in text preparation, feature generation, and creating, evaluating, and improving classification models. You will need to research NLP, and python functionalities if you aim to achieve **very good** marks and discover innovative techniques/methods.

Exploration, Preparation & Feature Generation

This section requires you to explore various aspects of your dataset and prepare the data for future sections. It is important you take time to carefully explore your data and make decisions on preparation or generation that make sense.

Preprocessing steps are essential to clean and standardize data before feature generation and enhance the quality of extracted features. Classification models that harness generated features may enable models to better understand and analyze data or to better learn patterns and relationships, compared to regular models.

Your task is to

- ➢ Explore and prepare your data.
  - ○ In this task, you could perform the necessary cleaning and pre-processing tasks, explore or try to understand and profile your data through various techniques (i.e. clustering, topic modelling, etc.).
- ➢ Generate new features from your data.
  - ○ You should have a good understanding of your data from above and can now experiment with feature generation. In this task you should consider what can be generated to improve your classification model.

Classification (Model Building and Evaluation)

It is **important** to try multiple variations of features/parameters in model building to achieve the best performance. Additionally, you should elaborate on the performance metrics you have used to evaluate your model and explain why they suit the available data.

Your task

- Experiment developing and evaluating classification models to find a model that has the best overall performance.
  - Once you find the best performing model, you should **only** show how you built and evaluated that specific one.
- Elaborate on the major tasks you have undertaken to improve the best-performing model and explain why the performance metrics suit the available data.

Submission

Your report should be delivered in an .ipynb file. A notebook template is provided to show how to structure your work. You **need to use** the template (Assignment_Template.ipynb) and strictly follow its format which is designed based on the provided Assignment rubric.

It can be useful that add some in-line comments (using #) next to your codes to explain it briefly.

You will get a better mark if your approach is **innovative**. This means no other student has applied it, or a few others have applied a similar approach with some differences. Therefore, it is **highly advised that you do not share your creative work with anyone else**. You can still discuss preliminary ideas and help each other, just remember your submission **must be your own work**.

To be done through Blackboard Assignment Submission, as indicated in Learn.UQ. The only acceptable submission format is .ipynb file. The file should be named in the format of YourStudentID.ipynb

You will only need to submit one .ipynb file and should use the provided Python template file.

Before submission:

- Ensure that your code can run without errors. If your code returns an error at any point, your assignment will only be marked up until the error, and the remainder of your code won't earn any marks. Example errors may include: Syntax issues or Name Errors.
- Make sure that all the important outputs are shown in your notebook. However, avoid showing trivial outputs. For example, you should remove codes randomly displaying the whole DataFrame, etc.

- Your marker will first look at your generated output as a reference **without running your notebook** (unless deemed necessary). Therefore, your significant outputs need to be generated, and the elaboration should be provided in the notebook, as shown in the template.