Machine Learning Canvas

	PREDICTIONS	OBJECTIVES	DATA
IDEA	**Context** Who will use the predictive system / who will be affected by it? Provide some background. Zameen.com, Pakistan's top property website, operates in various cities, enabling property transactions. The company aims to utilize this data to improve the accuracy of house price predictions and enhance the overall user experience on the platform.	**Value Proposition** What are we trying to do? E.g. spend less time on X, increase Y... We propose developing a machine learning model to automate the evaluation of house prices using diverse data sources. This solution will enhance pricing accuracy, build customer trust, optimize pricing strategies and maximize profits through precise, data-driven property valuations.	**Data Sources** Where do/can we get data from? (internal database, 3rd party API, etc.) The data for this project is sourced from a publicly available dataset on Kaggle. (https://www.kaggle.com/datasets/howisusmanali/house-price-prediction-zameencom-dataset)
SPECS	**Problem** Question to predict answers to (in plain English) What is the fair price for this property? Input (i.e. question "parameter") Area size, number of baths, bedrooms, etc. Possible outputs (i.e. "answers") Number from 0 to infinity (price in pakistani rupee) Type of problem (e.g. classification, regression, recommendation...) regression **Baseline** What is an alternative way of making predictions (e.g. manual rules based on feature values)? Real estate agents and agencies with human valuations	**Performance evaluation** Domain-specific / bottom-line metrics for monitoring performance in production - Cost Efficiency - Return on Investment - Operational Savings - Profit Margin Prediction accuracy metrics (e.g. MSE if regression; % accuracy, #FP for classification) - MAE less than 5% in train dataset and less than 10% in test dataset Offline performance evaluation method (e.g. cross-validation or simple training/test split) Simple training/test split	**Dataset** How do we collect data (inputs and outputs)? How many data points? Data provided by zameen.com, both inputs and outputs, in 5 batches totaling almost a year in length (sempling) **Features** Used to represent inputs and extracted from data sources above. Group by types and mention key features if too many to list all. - Location: location, city, latitude, longitude, province_name, location_id - Area: area, Area Type, Area Size, Area Category - Purpose of add: purpose - Agency/agent: agency, agent - Some characteristics of property: baths, bedrooms, property_type - Add related data: property_id, page_url, date_added
DEPLOYMENT	**Using predictions** When do we make predictions and how many? Every time a client/agent/company wants to know a fair price for a property What is the time constraint for making those predictions? Less than 1 second for 1 prediction How do we use predictions and confidence values? We compare it with the offered price, add there our personal opinion (related on the photos of the property, something else) and decide whether it is a normal price, it should be lower or higher.		**Learning predictive models** When do we create/update models? With which data / how much? When each new batch appears (every 2 months), you need to do a retraining on the new data. The data are provided by Zameen.com and have an average size of about 30000-40000 rows What is the time constraint for creating a model? 24 hours max Criteria for deploying model (e.g. minimum performance value — absolute, relative to baseline or to previous model) MAE at most 15% on unseen data Note: in some periods of volatility, prices may be less predictable, then the outcome may be slightly worse

Reset Form

Machine Learning Canvas v0.1