Machine Learning Round

Duration: 40 Minutes
Difficulty: Medium
Domains: Finance

Problem

Redfin is a real estate platform where users can sell and buy houses at market prices. Design a machine learning system that optimizes price for home sellers.

#1 - How would you predict that a house will sell in 30 days?

#2 - How would you optimize the price to maximize the probability that a house sells in 30 days and sale price?

Solution

> **#1** - How would you predict that a house will sell in 30 days?

[Candidate] I will start with an assumption that this problem involves a classification model. The outcome is binary - a house sells in 30 days or not. Also, I assume that I can use data such as price, amenities, and neighborhoods.

[Interviewer] Yes, that is correct. Given that the problem is classification, how would you create your label?

[Candidate] I will create a label using the listing and selling timestamps. If the subtraction of selling timestamp from listing timestamp is less than or equal to 30 days, assign "1" to the house sold. A label of "0" is assigned to unsold houses.

[Interviewer] Do you see an issue if you used houses listed in the most recent 30 days?

[Candidate] Yes, houses listed recently are not observed during the entire 30-day period for labeling. Such label issue is known as right-censoring. To address this issue, houses listed in the most recent 30 days are removed.

[Interviewer] Let's discuss feature engineering. Given the variables below, explain your approach to feature engineering:

1. Home address
2. Count of bedrooms and bathrooms
3. Square footage
4. Sale price
5. Listing date
6. Sold date

[Candidate] In the variable, home address, can I presume that this field contains street address, neighborhood, state and zip code?

[Interviewer] Yes.

[Candidate] Using a string parser, I would decompose home address string to create variables including street name, neighborhood, state, and zip code.

[Interviewer] How would you handle high-cardinality of those categorical features you mentioned?

[Candidate] One-hot encoding the categorical fields will not be appropriate given sparsity of the feature set, causing a model to overfit. Hence, the features require conversion to continuous values. I would apply mean encoding on the categorical variables. For each value in the categorical variable, compute the mean of sale price and encode it with the mean. For instance, for zip code XYZ, average the sale price and encode the average in the zip code.

[Interviewer] Suppose you have your feature set, how would you build your model?

[Candidate] I would split the data into train, validation and test based on listing date. Compute the encoded values in the train set and replace the values for the rest of the splits. This approach will prevent data leakage.

Next, I would start with a simple model either a regularized logistic regression model or random forest. Evaluate model performance using AUC.

[Interviewer] How would you assess that your model is overfitting?

[Candidate] I would assess the AUC's of train, validation and test splits. A large discrepancy, let's say 0.10+ difference of AUC between valid and test suggests that a model fails to generalize given overfitting.

Interviewer Feedback: The candidate demonstrated strong performance in communication and grasp in machine learning fundamentals. First, she correctly identified that the problem type is classification. She then proposed a sound approach to label creation. Handling right-censored data with a deletion of most recent records made sense.

Next, when asked about feature engineering, she proposed deriving additional fields from home address. One suggestion for her is to explain how the granular fields could benefit the model. What information does geographical details carry about the likelihood of a house sold in 30 days? I presume that, depending on areas, the average number of days a house remains unsold varies across areas.

When asked about overfitting, she avoided a common pitfall in feature engineering which is one-hot encoding of a high-cardinality variable. Instead, she proposed a sound technique involving numerical encoding.

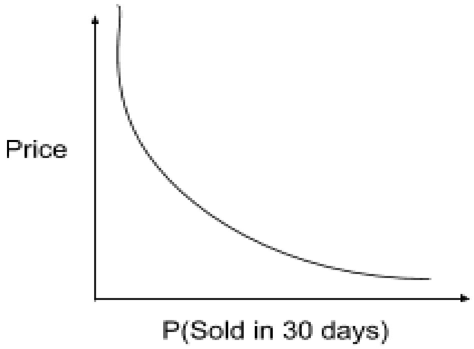
She finalized with sound suggestions on data splitting for training and evaluation, classic statistical models and assessment of overfitting in model performance. Overall, she explained her approach clearly and demonstrated a strong foundation in machine learning design.

> **#2 -** How would you optimize the price to maximize the probability that a house sells in 30 days and sale price?

[Candidate] This is a tricky problem involving optimization of price. I'm assuming that I will need to use the classification model proposed in the previous problem. I will need to use the probability, not the class, output of the model.

[Interviewer] Yes, you probably need to.

[Candidate] I presume that there is an inverse relationship between price and the likelihood of a house sold in 30 days. For instance, if the price of a home decreases, the likelihood increases.



[Interviewer] That's a fair assumption.

[Candidate] I think I may have a solution. I know that one area actively researched in machine learning is interpretability of non-linear models such as GBM and neural networks. One method to interpret the relationship between target and predictor variables is the partial dependence plot.

[Interviewer] Okay, how would you use this method?

[Candidate] Since my model incorporates the price as a predictor, the relationship between price and probability is expressed in the following:

$$P(Sold \mid X_1 = Price, X_2, ..., X_p)$$

For a home, holding all other variables constant, I observe the relationship between various price points and the probability of selling it.

[Interviewer] Now how would you choose an optimal price on probability and profit?

[Candidate] I would pick the one that maximizes the probability.

[Interviewer] Are you sure? Wouldn't your profit approach $0 given that, to maximize the probability, the price approaches 0?

[Candidate] Actually you are right. Let me correct my thinking. In fact, I would use a probability concept called expected value.

$$Expected\ Value_k = Price_k * P(Sold)_k$$

For each price point, *k*, I would compute the expected value of the profit. The price point that maximizes the expected value should be recommended.

Interviewer Feedback: The candidate provided a correct solution to the problem and explained her approach clearly.

First, she recognized that the probability of a home sold in 30 days has an inverse relationship with price. Her inkling suggests that she grasps business-sense required for the machine learning role.

To simulate the relationship, she suggested using partial dependence plot. Her approach was correct; the partial dependence plot explains the relationship between a target variable and predictor.

When asked about price optimization, her initial response of setting the price that maximizes the probability was incorrect. Given the inverse relationship between the probability and price, the highest probability is at 0. Hence, the profit becomes 0, which fails to optimize not just the probability but also profit.

Nonetheless, she corrected herself with the correct approach maximizing the expected value of a home sale given price multiplied with the corresponding probability.

Final Assessment

In the statistics section, a candidate is assessed based on correctness and soundness of statistical methodology, business sense and communication. For each dimension the candidate is rated in the following scale: (5) superior, (4) good, (3) adequate, (2) marginal, (1) not competent.

Assessments	Rating	Comments
Machine Learning Methodology	5	The candidate demonstrated strong skills in machine learning as she offered sound solutions without haste. For instance, when asked her about avoiding overfitting on a feature with high-cardinality, she proposed numerical encoding, which could work. In addition, she proposed AUC to evaluate the classification model, which is a sound approach. Lastly, in the final problem involving price optimization, her approach to calculating expected value is a valid approach.
Product Sense	5	The second question assessed her product sense. She understood how to relate price and demand to the the modeling problem. She understood that she could leverage the partial dependence plot to calculate the probability of a home sold in 30 days with respect to price. Although she initially made a minor error on price optimization, she eventually corrected herself with expected value for the profit maximization.
Communication	5	The candidate’s demonstrated strong communication skills with clear explanations. She grasped the problems quickly and expressed her approaches clearly.