


# Home-depot product search relevance

Yixin Zhang, Yuantai Xie | 2020 Aug 21



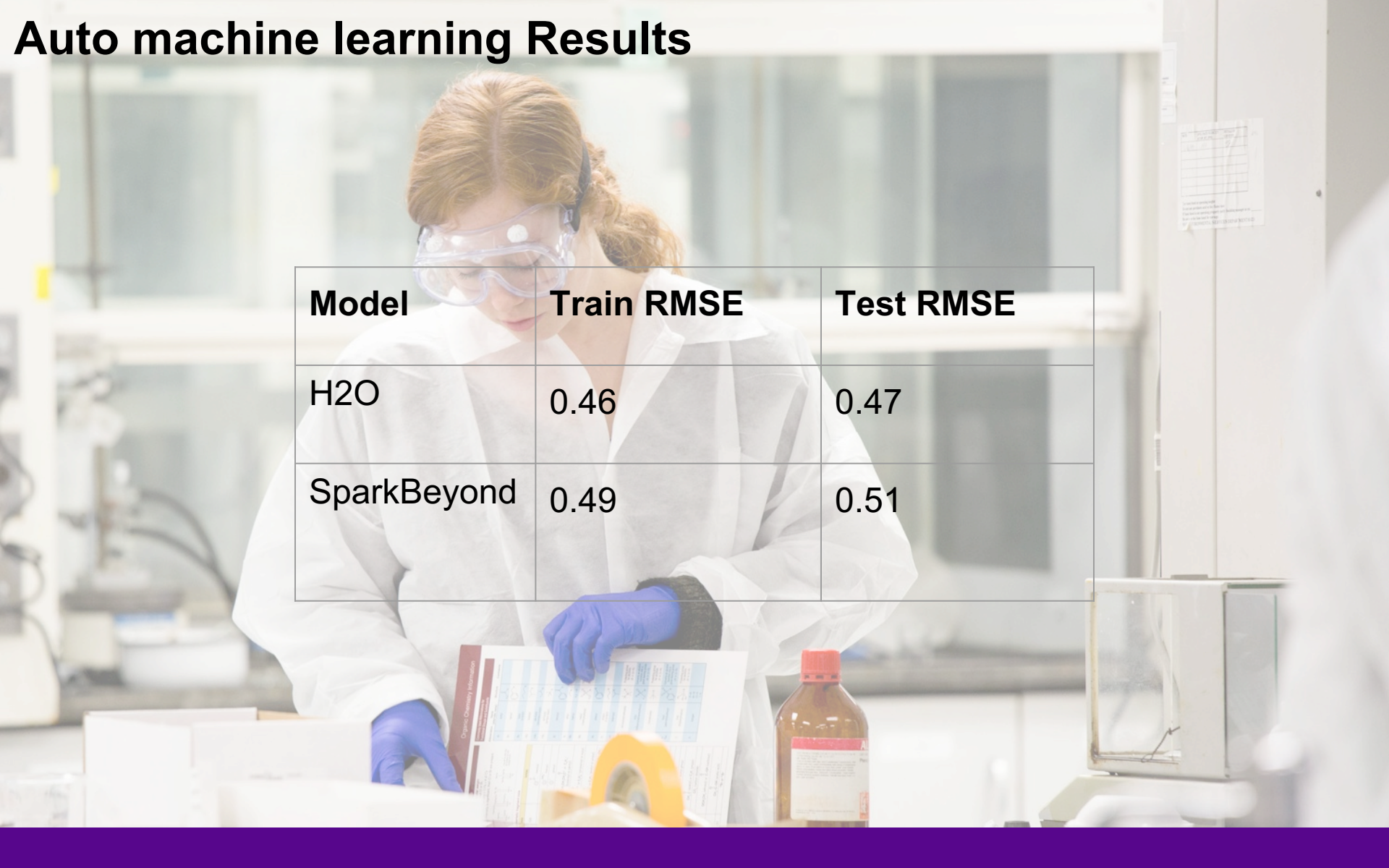
NYU

# Results

A background image of a female scientist with reddish-brown hair, wearing a white lab coat, safety goggles, and blue gloves. She is working at a lab bench, looking down at some papers. On the bench, there is a brown bottle, a yellow container, and some other lab equipment. The image is slightly blurred, focusing on the scientist and the overlaid table.

| Model    | Details                              | NDCG metric | Train RMSE | Test RMSE |
|----------|--------------------------------------|-------------|------------|-----------|
| GBM      | {'max_depth': 6, 'n_estimators': 40} | 0.988       | 0.45       | 0.47      |
| XGBoost  | {'max_depth': 4, 'n_estimators': 32} | 0.987       | 0.46       | 0.48      |
| LightGBM | {'num_leaves': 15}                   | 0.986       | 0.52       | 0.53      |

# Auto machine learning Results

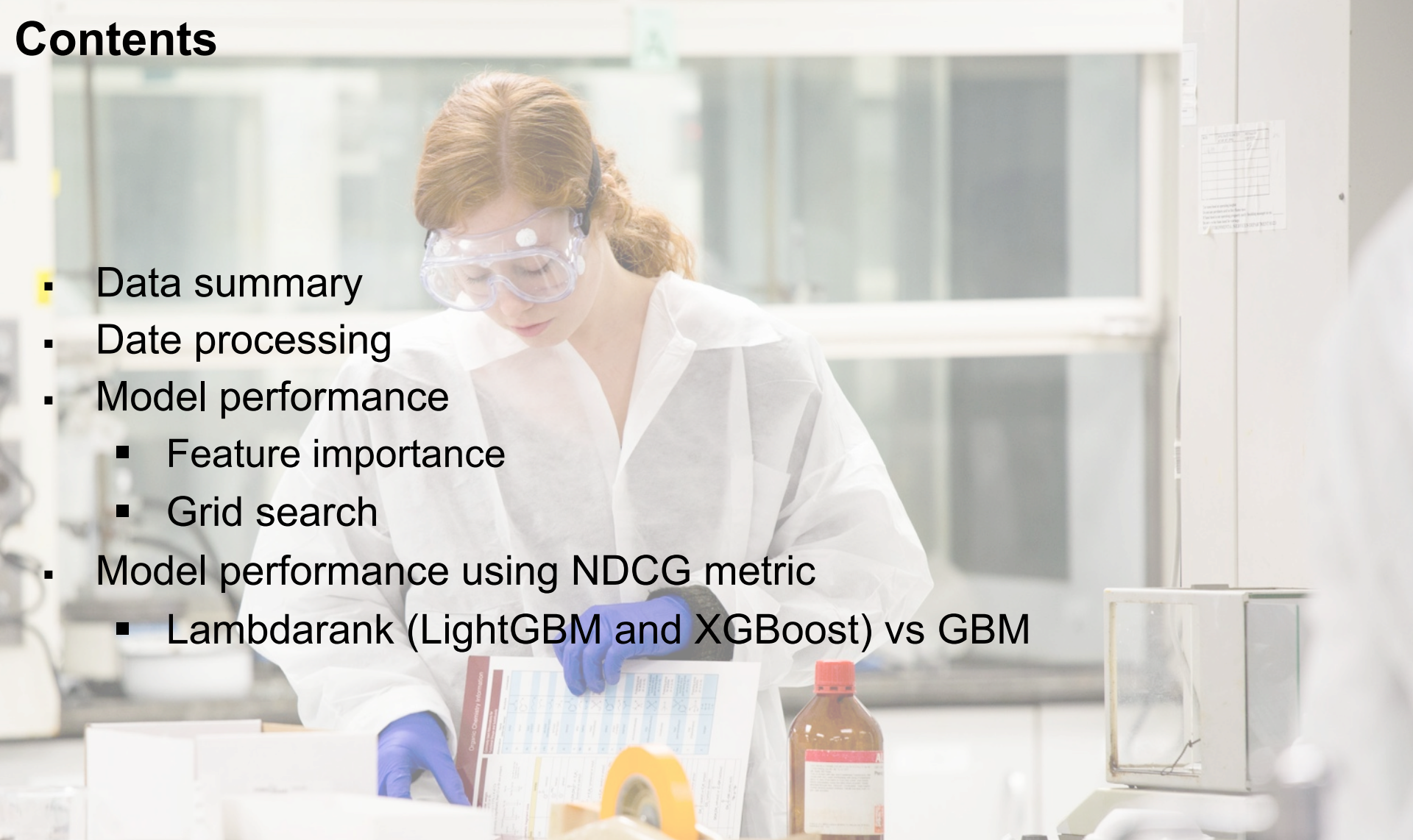
A female scientist with reddish-brown hair, wearing a white lab coat, safety goggles, and blue nitrile gloves, is working in a laboratory. She is looking down at a document or piece of equipment. In the background, there are various laboratory instruments and equipment. A table with machine learning results is overlaid on the image.

| Model       | Train RMSE | Test RMSE |
|-------------|------------|-----------|
| H2O         | 0.46       | 0.47      |
| SparkBeyond | 0.49       | 0.51      |



# Contents

- Data summary
- Data processing
- Model performance
  - Feature importance
  - Grid search
- Model performance using NDCG metric
  - Lambdarank (LightGBM and XGBoost) vs GBM



# Data summary

- Predict the *relevance* of search results with products

|   | id | product_uid | product_title                                     | search_term        | relevance |
|---|----|-------------|---|--------------------|-----------|
| 0 | 2  | 100001      | Simpson Strong-Tie 12-Gauge Angle                 | angle bracket      | 3.00      |
| 1 | 3  | 100001      | Simpson Strong-Tie 12-Gauge Angle                 | l bracket          | 2.50      |
| 2 | 9  | 100002      | BEHR Premium Textured DeckOver 1-gal. #SC-141 ... | deck over          | 3.00      |
| 3 | 16 | 100005      | Delta Vero 1-Handle Shower Only Faucet Trim Ki... | rain shower head   | 2.33      |
| 4 | 17 | 100005      | Delta Vero 1-Handle Shower Only Faucet Trim Ki... | shower only faucet | 2.67      |

(74067, 5)

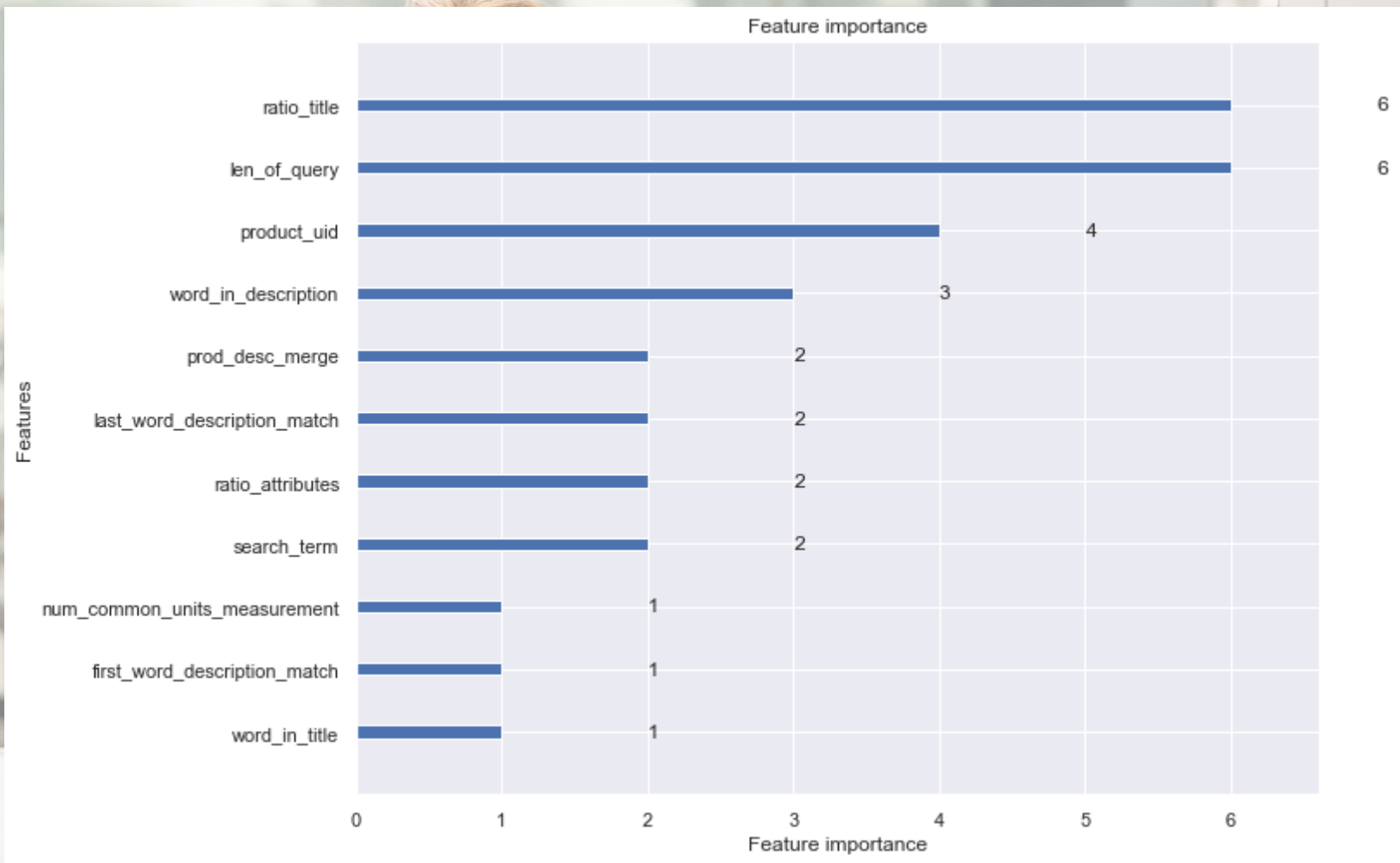
# Data processing

- 1) Spell checking
  - Use a pre-defined dictionary of spelling mistakes
  - 3,400 pairs
- 2) Stemming
  - Perform stemming function on variables *search\_term*, *product\_title*, *product\_description*, and *product\_attributes*
- 3) Create 16 new features
- 4) Calculate TF-IDF and do feature reduction

## VARIABLE IMPORTANCE

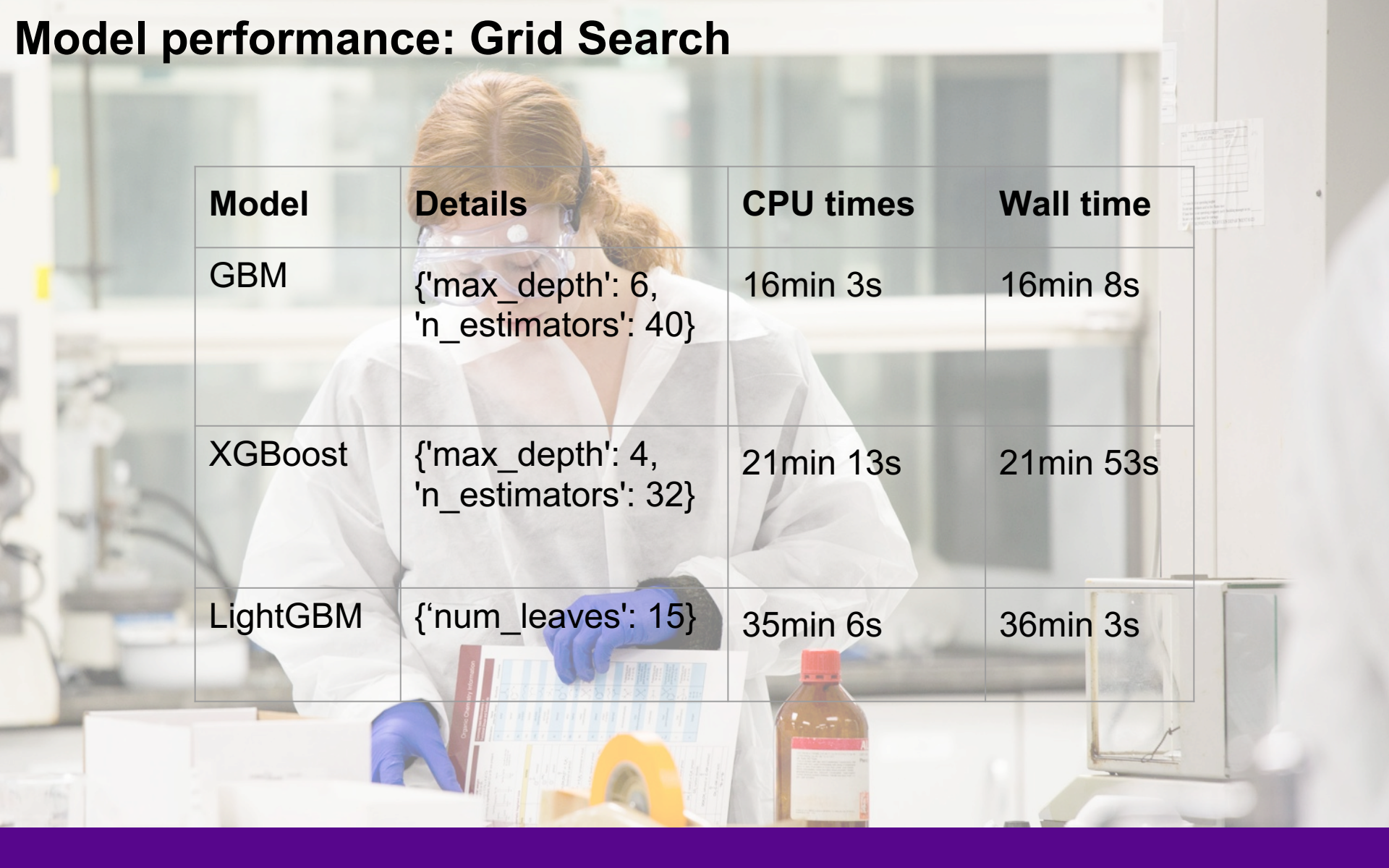
|                             |      |
|-----------------------------|------|
| 6_CVTE:search_term.0        | 1.00 |
| 7_TxtTE:search_term.0       | 0.22 |
| 0_product_uid               | 0.02 |
| 15_NumToCatTE:product_uid.0 | 0.02 |
| 16_NumToCatTE:product_uid.0 | 0.01 |
| 17_Txt:search_term.0        | 0.01 |
| 17_Txt:search_term.8        | 0.01 |
| 17_Txt:search_term.1        | 0.01 |
| 17_Txt:search_term.48       | 0.00 |
| 17_Txt:search_term.7        | 0.00 |
| 17_Txt:search_term.45       | 0.00 |
| 17_Txt:search_term.49       | 0.00 |
| 17_Txt:search_term.18       | 0.00 |
| 17_Txt:search_term.28       | 0.00 |

# Model performance: Feature importance





# Model performance: Grid Search

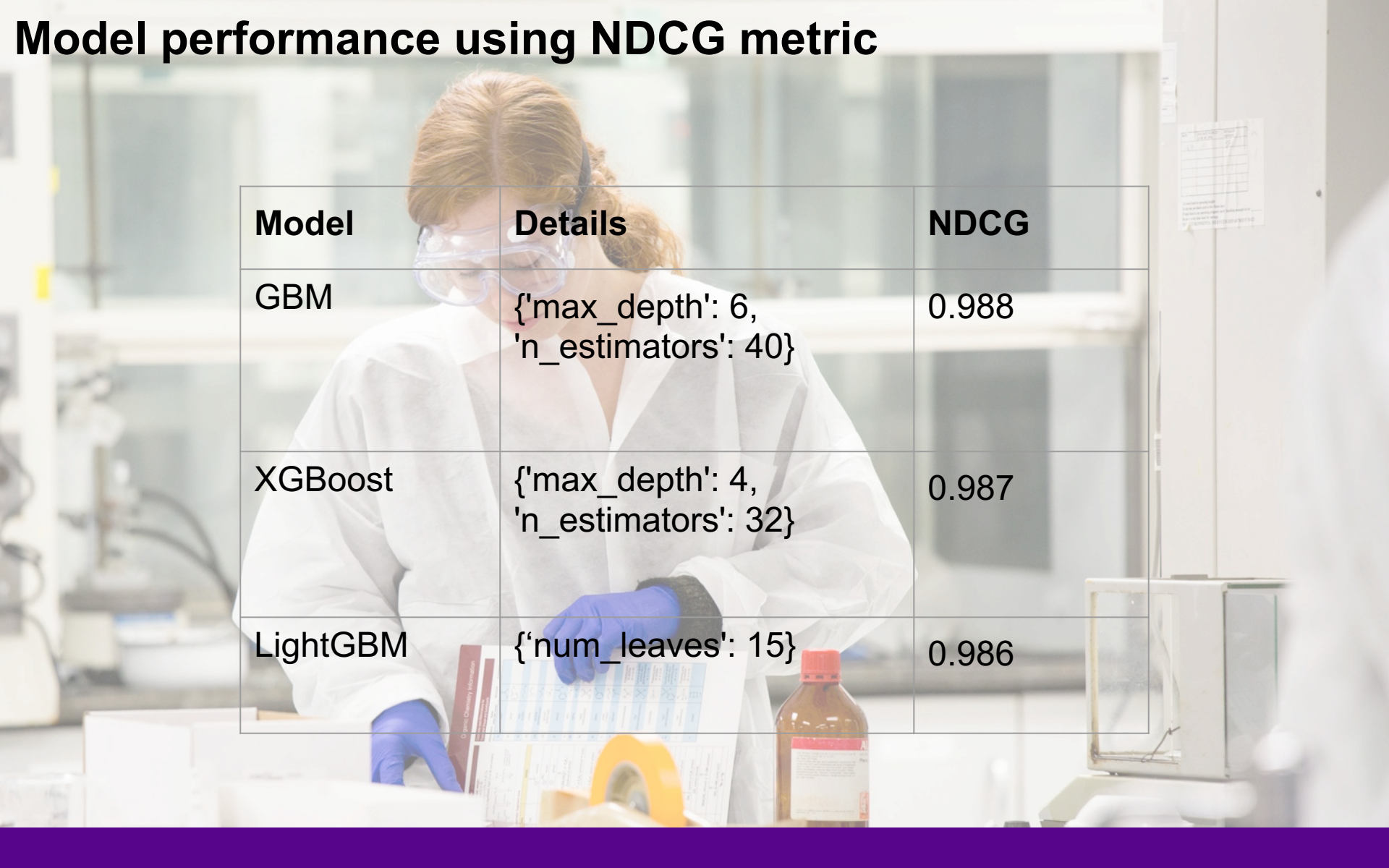


| Model    | Details                                 | CPU times | Wall time |
|----------|---|-----------|-----------|
| GBM      | {'max_depth': 6,<br>'n_estimators': 40} | 16min 3s  | 16min 8s  |
| XGBoost  | {'max_depth': 4,<br>'n_estimators': 32} | 21min 13s | 21min 53s |
| LightGBM | {'num_leaves': 15}                      | 35min 6s  | 36min 3s  |

# Intro and code of NDCG metric

```
01. """
02. This function calculates ndcg for multiple queries dataset
03. y_true : array, shape = [n_samples]
04.         Ground truth (true relevance labels)
05. y_score : array, shape = [n_samples]
06.         Predicted scores
07. k:      int, optional
08.         Only consider the highest k scores in the ranking. If None, use all outputs.
09. group: array, shape = [n_groups]
10.        each element denotes how many items are there in each group
11. assume all queries have equal weights
12. """
13.
14. def ndcg_score(y_true, y_score, group, k = None):
15.     avg_ndcg = 0
16.     index = 0 #next row to be calculated
17.     count = 0 #number of groups which can provide information
18.     for i in range(0, len(group)):
19.         cur_true = y_true[index: index+group[i]-1]
20.         cur_score = y_score[index: index+group[i]-1]
21.         index += group[i]
22.
23.         idcg = dcg_score(cur_true, cur_true, k)
24.         # when ground truth is equal to 0, we abandon that group which provides no information
25.         if idcg == 0:
26.             continue
27.
28.         cur_ndcg = dcg_score(cur_true, cur_score, k)/idcg
29.         avg_ndcg = cur_ndcg * 1/(count+1) + avg_ndcg * count/(count+1)
30.         count += 1
31.
32.     return avg_ndcg
33.
34. def dcg_score(y_true, y_score, k = None):
35.     order = np.argsort(y_score)[::-1]
36.     y_true = np.take(y_true, order[:k])
37.
38.     gain = 2 ** y_true - 1
39.
40.     discounts = np.log2(np.arange(len(y_true)) + 2)
41.     return np.sum(gain / discounts)
```

# Model performance using NDCG metric

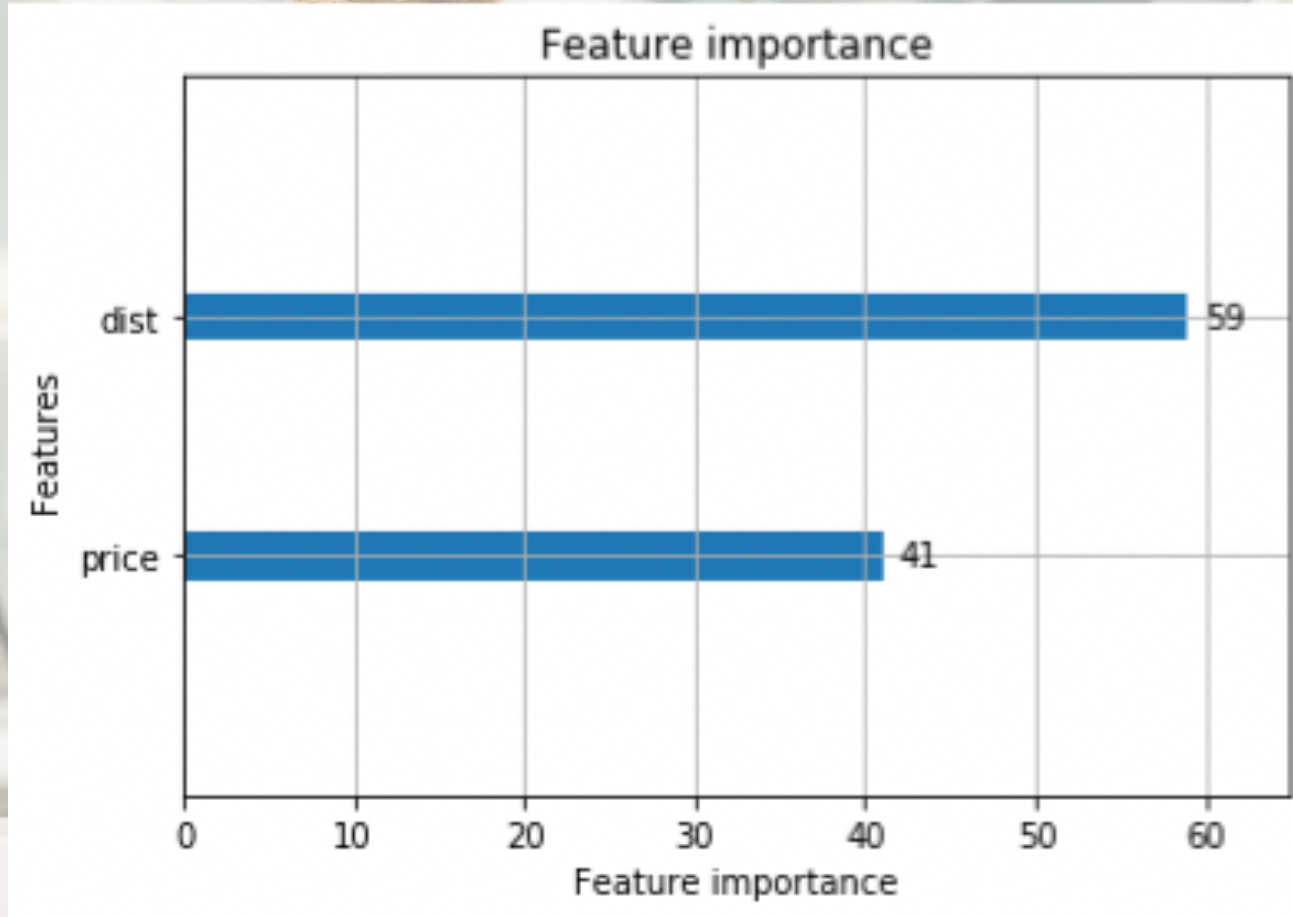


A scientist with red hair, wearing a white lab coat, safety goggles, and blue gloves, is working in a laboratory. She is looking down at a document or equipment. In the background, there are various lab equipment and a window. A table is overlaid on the image, showing the performance of three machine learning models using the NDCG metric.

| Model    | Details   | NDCG  |
|----------|---|-------|
| GBM      | <code>{'max_depth': 6, 'n_estimators': 40}</code> | 0.988 |
| XGBoost  | <code>{'max_depth': 4, 'n_estimators': 32}</code> | 0.987 |
| LightGBM | <code>{'num_leaves': 15}</code>                   | 0.986 |



# LGBMRanker model running



***Thanks for listening!***