DTSA 5509 Project

March 25, 2024

Digital Green Crop Yield Estimate Challenge

ZiNDi

- Zindi is African version of Kaggle
- Zindi hosts the largest community of African data scientists, working to solve the world's most pressing challenges using machine learning and AI

Digital Green

- Digital Green is the main sponsor for this challenge
- The objective of this challenge is to create a machine learning solution to predict the crop yield per acre of rice or wheat crops in India. The goal is to empower these farmers and break the cycle of poverty and malnutrition

Details for this challenge: https://zindi.africa/competitions/digital-green-crop-yield-estimate-challenge

Data Overview

- This comprehensive survey was carried out across various districts in India, aiming to gather detailed insights into factors influencing rice crop yields.
- Key focus areas included fertilizer usage, seeding quantities, land preparation techniques, and irrigation methods, among others
- The survey resulted in a rich dataset containing over 5,000 data points, each described by more than 40 distinct features.
- These features encapsulate a wide range of agricultural practices and inputs potentially affecting rice yield.
- The primary goal is to leverage this dataset to develop predictive models that can accurately forecast rice crop yields based on the myriad factors recorded.

Problem Statement

We need to predict the yield from 42 other features from the dataset.

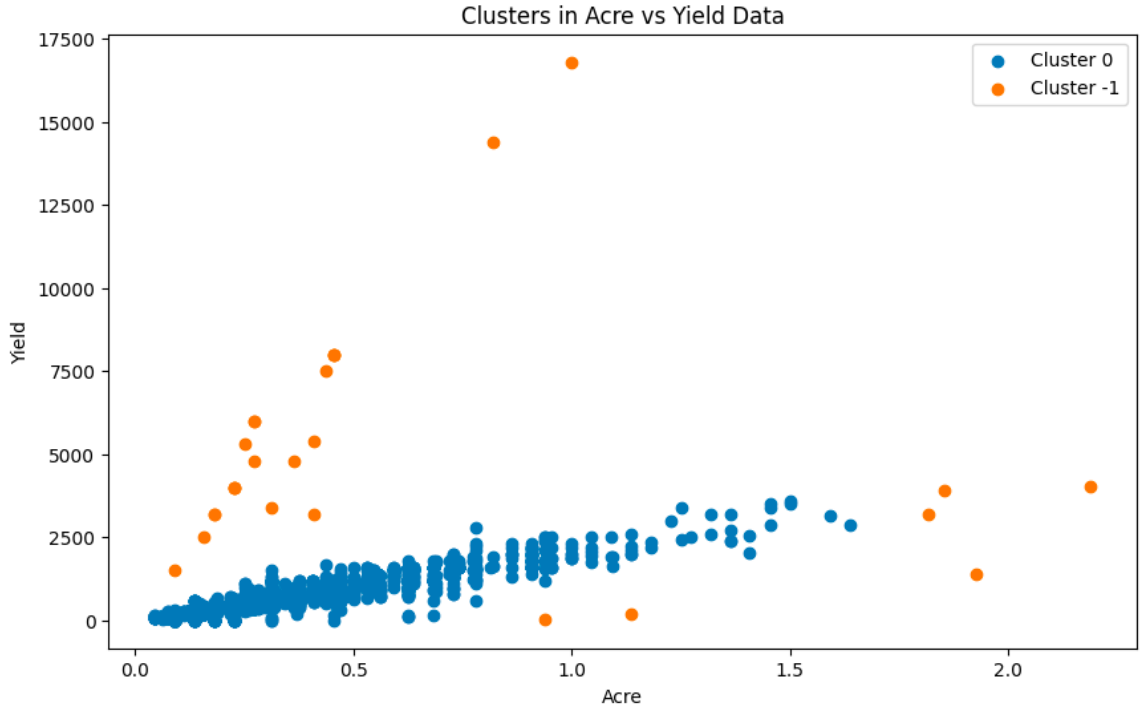
The evaluation metric is Root Mean Square Error.

Reach as low as possible RMSE to score higher on the leaderboard.

Most Important feature

Acres

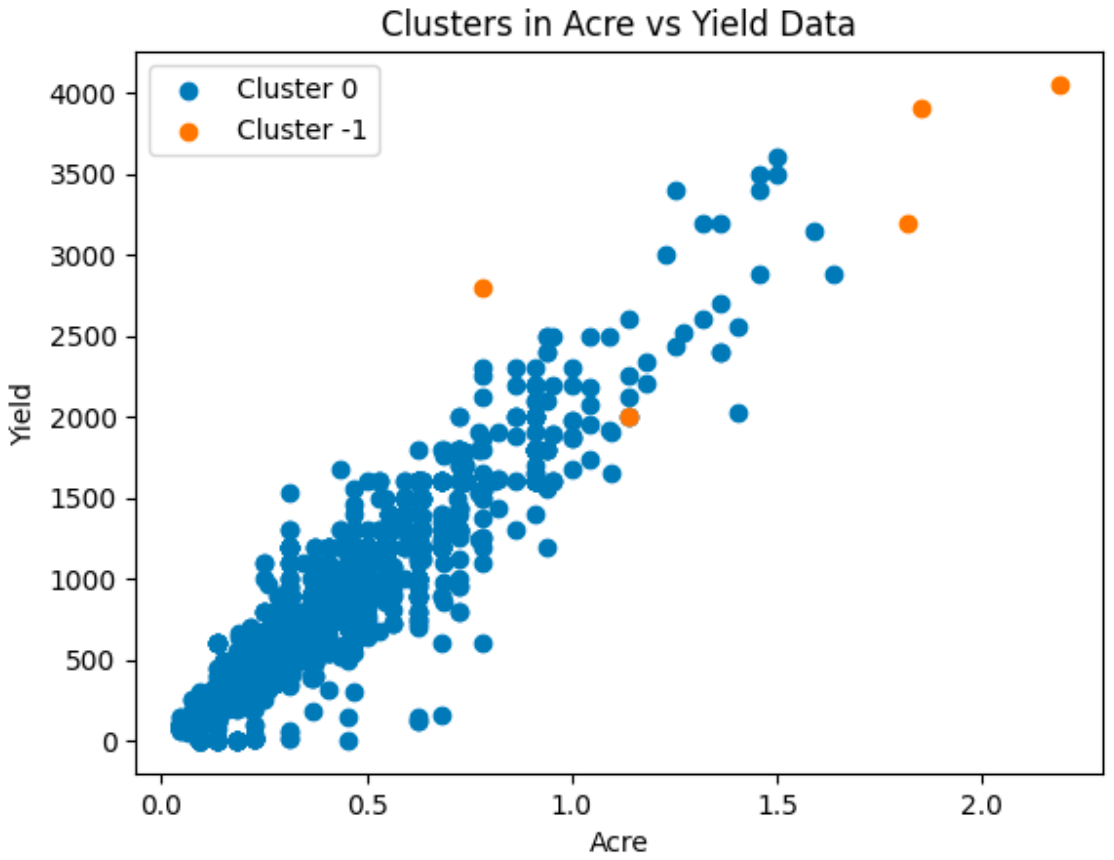
- The area of the land has the largest impact on the yield.
- Looks like a linear regression relationship
- Two clusters are found using Density-Based Spatial Clustering of Applications with Noise (DBSCAN)



Real World Problem

Data is not always clean!

- Some outliers in the dataset cause the other cluster.
- Remove and refit some data points for training.



Outliers in testing data

- The largest problem for all the challengers.
- RMSE will be impact by a single outlier very easily. One outlier can increase about 100-200 RMSE from actual value.
- Challengers need to find the outliers in testing data to achieve higher score on public leaderboard

My Solution

Public leaderboard Tunning

Fun part that makes this model performs well in public leaderborad.
I was able to find two outliers in the public leaderboard that was hard to get using this method.
First one was the common outlier 'ID_PMSOXFT4FYDW', this model will predict it as 787, which is aboiut 1/10 of the real value.
The second one is 'ID_BI4VNVU7JAXF', which is also strange, my predicted value was 1/2 of the true value.

```
[21]
# Public leaderboard tunning
print(submission.loc[submission[id] == 'ID_PMSOXFT4FYDW', label_col])
submission.loc[submission[id] == 'ID_PMSOXFT4FYDW', 'Yield'] = 8000
print(submission.loc[submission[id] == 'ID_BI4VNVU7JAXF', label_col])
submission.loc[submission[id] == 'ID_BI4VNVU7JAXF', 'Yield'] = 3200
```

```
373    787.0
Name: Yield, dtype: float64
1118    1520.0
Name: Yield, dtype: float64
```

- Data probing, trial and error test to find the potential outliers.
- Found 2 outliers as shown in the picture.
- Got No.3 in Public Leaderboard by manual changing the values.
- However public leaderboard does not affect final score, we still need to get better score on private leaderboard.

Different Approaches

There are discussions between participants in the discussion board of this challenge.

Some suggest predicting the error is much appropriate, so they prone to use more complicate and overfitting models

Others suggest to drop all outliers, use simpler models and ignore the potential outliers in final test.

Models

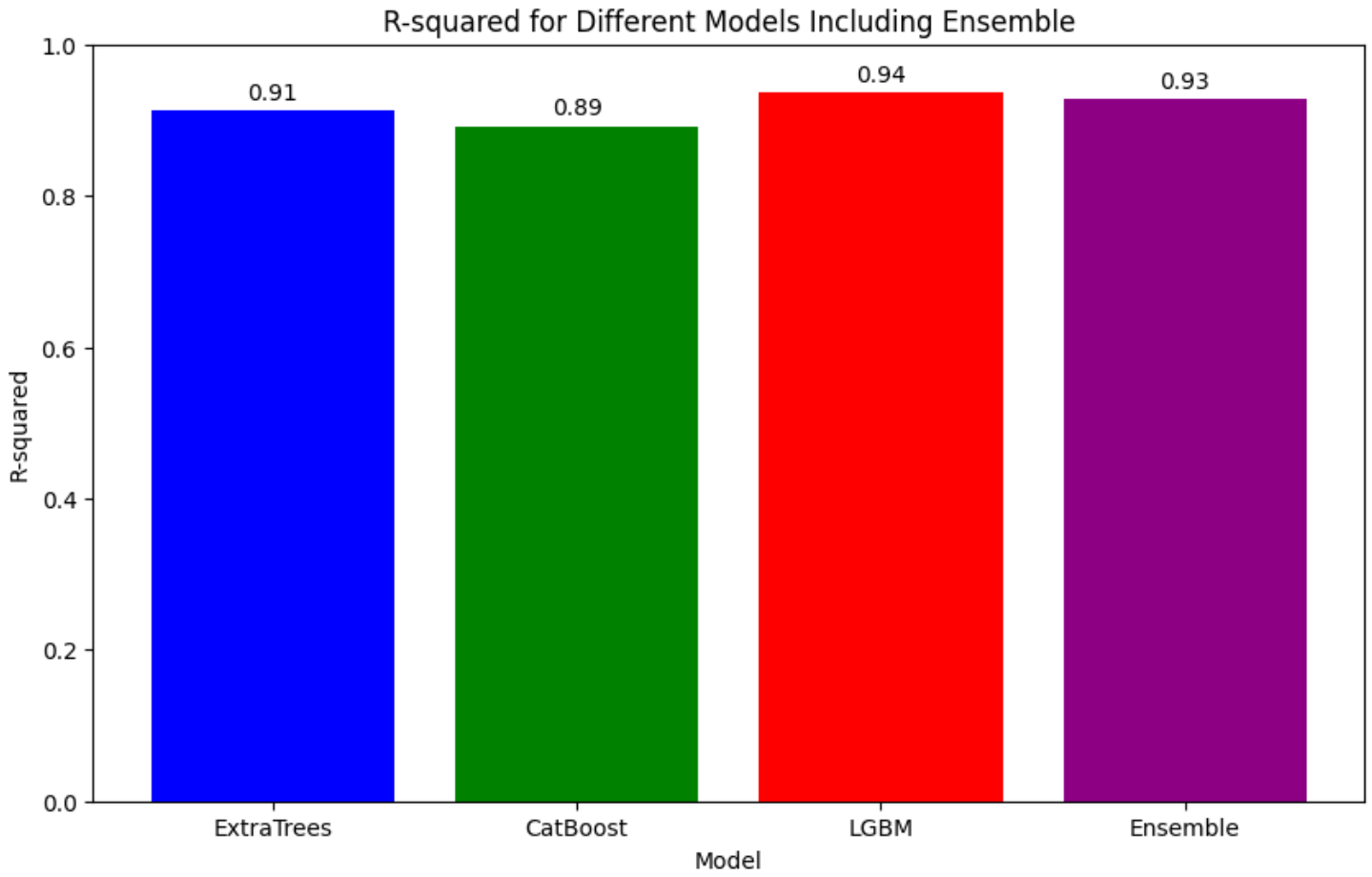
By limitation of computing power, I choose to use simpler models.

Three models

- Extra Trees
- Catboost
- Lightgbm

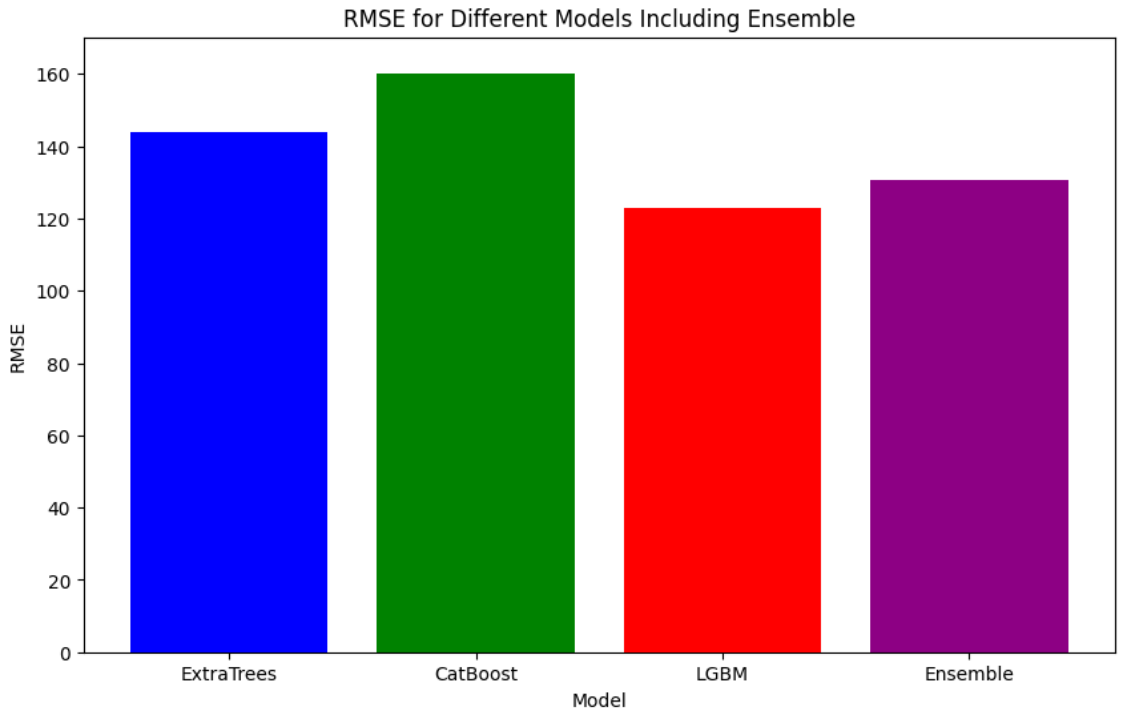
According the rule, I can only ensemble three models. So I ensemble the three models by taking average of the three.

Model Performance (R^2)



Model Performance (RMSE)

- LightGBM seems have higher score. (The lower RMSE the better)
- Ensemble model is more robust.
- Final score favors Ensemble model.



Final Score

20		cliff003	109.5486094	120.6283946	Download	4 months ago	83

20th among 678 challengers

The running time of my model was under 9 seconds!
Simple modeling works well for this problem.

Conclusion & Final thoughts

- This was a very tense work, lots of work done were before modeling.
- Real-world data need to be cleaned carefully.
- Feature engineering is not working for my models, more added features gave worse results.
- Also complicated and time-consuming models (more than 1 day to run) did not improve the score.
- Data challenges are not very practical, we only care about final scores instead of reasoning. Some of the code was only for the improvement of the score such as round the result to integer.

More..

- Domain knowledge is very important in machine learning. Lack of agricultural experience makes me difficult to understand some features.
- Need to collaborate with domain expert, communication is key to do data science jobs.

- A team from Oxford University (3rd on the leaderboard) has provide comprehensive insights and solution for this challenge.
- Their team GitHub:

https://github.com/rapsoj/crop-yield-estimate/tree/main

Thank You!