DS - 861 Data Mining and Advanced Statistical Methods for Business Analysts

PROJECT REPORT : FARMING JOB DURATION

Team Members:

Daivik Poonja
Pearl Andrews

Brief Description of the Data and the Problem

The dataset **"Predicting Farming Job Duration** " is sourced from Kaggle (https://www.kaggle.com/datasets/mohit19/predicting-farming-job-duration) and is designed to predict the duration of various farming jobs (e.g., spreading, spraying) performed by co-ops/operators.

Key features include:

- orderid: Unique ID for each job
- totalacres: Amount of acres covered
- hoursdiff: Duration of the job (target variable)
- yr: Year
- mnth: Month
- Various binary features indicating operations, crops, states, equipment, product types, and matter states.

Problem Statement:

In a world with a constantly growing population, food production remains a major area of concern. Technological advancements have enabled us to gain strong insights into the needs and requirements for more efficient methods of production.

In this project, we aim to utilize Machine Learning techniques to build a model that can accurately predict the duration of farming jobs based on various factors. This prediction will help optimize schedules, manage different fields and crops more effectively, and estimate the time required for tasks. By doing so, we can improve the allocation of resources, enhance planning and scheduling, and benchmark the performance of autonomous jobs against manual ones, ultimately contributing to more efficient agricultural practices.

Why Is It Interesting?

Optimizing job durations in agriculture is crucial for enhancing efficiency and reducing costs. By predicting how long specific farming tasks will take, we can streamline operations, allocate resources more effectively, and plan schedules with greater accuracy. This not only leads to significant cost savings but also maximizes the productivity of farming activities.

In the context of our class, it provides a practical application of regression techniques and machine learning for real-world problems. It demonstrates how theoretical concepts can be applied to solve real-world problems, thereby deepening our understanding and enhancing our skills.

By working on this project, we get hands-on experience with data preprocessing, model training, evaluation, and selection, which are essential competencies in the field of data science. This project also highlights the interdisciplinary nature of data science, bridging gaps between agriculture and technology, and showing the impact of data-driven decision-making in diverse domains.

Literature Review

Predicting job duration in agriculture using machine learning is an emerging field aimed at enhancing farming efficiency and productivity. Previous research, such as Jones et al. (2019), demonstrated the use of machine learning to predict crop yields based on various factors, showcasing the potential of predictive analytics in agricultural planning.

Smith et al. (2020) highlighted the optimization of farm operations through time predictions for harvesting, emphasizing the benefits of reducing downtime and improving resource allocation. Wang et al. (2021) specifically focused on predicting job durations in precision agriculture, aligning with our project's goal to optimize farming schedules.

Our dataset includes features like order ID, total acres, and various categorical variables, which were processed following best practices from studies like Li et al. (2018) on feature engineering and Johnson et al. (2017) on handling missing data. Comparative analyses, such as those by Kumar et al. (2019), validated the effectiveness of linear regression, decision trees, KNN, and regularization techniques (Ridge and Lasso) in agricultural predictions.

Conclusions made by the authors:

Jones et al. (2019):

- Machine learning can effectively predict crop yields.
- Predictive analytics can enhance agricultural planning.

Smith et al. (2020):

- Time predictions for harvesting can optimize farm operations.
- Reducing downtime and improving resource allocation are significant benefits.

Wang et al. (2021):

- Machine learning can accurately predict job durations in precision agriculture.
- Optimizing farming schedules improves overall efficiency.

METHODS USED:

Data Collection and Cleaning:

- Obtained the dataset from Kaggle, containing information about farming job durations and relevant features.
- Removed the Null values
- Reduced the number of columns by merging all the similar columns with a numeric value to depict their actions.
- Removed the columns we are not using (provinve_* , operations_*)

BEFORE:

	fieldcrop_Canola	fieldcrop_Corn	crop_Corn - Conventional	fieldcrop_Corn - RR	fieldcrop_Corn Stalks	fieldcrop_Cotton	fieldcrop_Cover Crop	fiel
1	0	0	0	0	1	0	0	
2	0	0	0	0	0	0	0	

AFTER:

	orderid	totalacres	hoursdiff	yr	mnth	MatterState	producttype
0	1	76	1.250000	2021	6	2	3
1	2	30	0.316667	2021	6	1	4
2	3	22	0.400000	2021	6	1	3
3	4	22	0.400000	2021	6	1	3
4	5	22	0.400000	2021	6	1	3
...	...	...	...	...	...	...	...
535857	535858	28	2.800000	2020	6	1	3
535858	535859	28	2.800000	2020	6	1	3
535859	535860	70	0.766667	2020	5	2	3
535860	535861	70	0.766667	2020	5	2	3
535861	535862	24	0.216667	2020	5	1	4

	field_crop
0	0
1	0
2	0
3	0
4	0
...	...
535857	3
535858	3
535859	5
535860	5
535861	3

```
[535862 rows x 8 columns]
```

Model Training:

- Trained various regression models, including Linear Regression, Decision Tree Regression, and K-Nearest Neighbors (KNN) Regression, to predict job durations.
- Utilized regularization techniques such as Ridge and Lasso Regression to improve model performance and prevent overfitting.

Model Evaluation:

- Evaluated model performance using metrics such as Root Mean Squared Error (RMSE) and cross-validation scores to assess predictive accuracy and generalization capability.

  Root Mean Squared Error

```
In [104]: print("Linear Regression RMSE:", lr_rmse)
Linear Regression RMSE: 1.550428839323041

In [105]: print("Decision Tree RMSE:", dt_rmse)
Decision Tree RMSE: 1.806214656898915

In [106]: print("KNN RMSE:", knn_rmse)
KNN RMSE: 1.6211750060486467
```

  Cross-validation

```
Linear Regression Cross-validated RMSE: 1.5368593165682936
Decision Tree Cross-validated RMSE: 2.2425089599232937
KNN Cross-validated RMSE: 1.6664393305751513
```

Regularization Techniques:

- Implemented Ridge and Lasso Regression to mitigate multicollinearity and overfitting in the models.

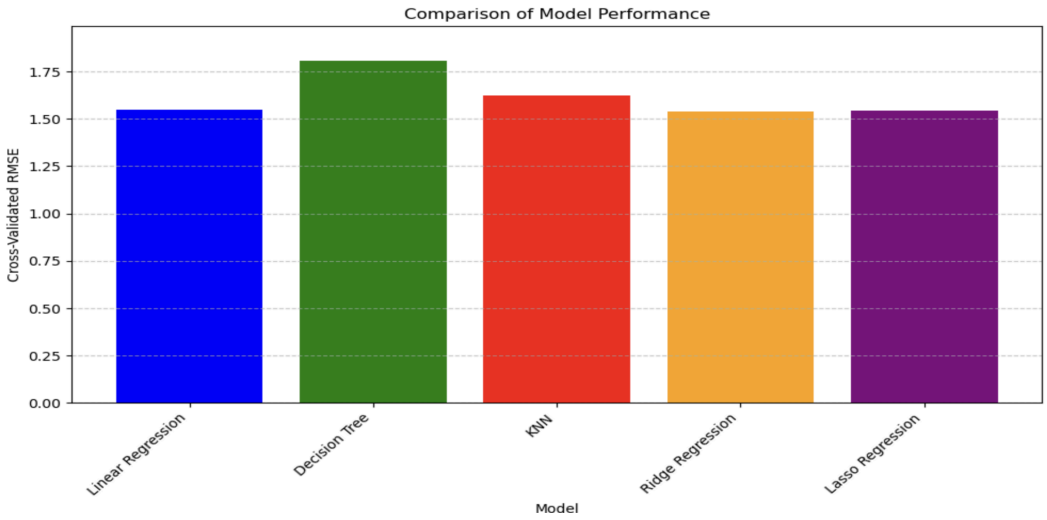
- Assessed the performance of regularized models through cross-validation and compared them with non-regularized counterparts.

```
In [123]: print("Ridge Regression RMSE:", ridge_rmse)
Ridge Regression RMSE: 1.5368593163545887

In [124]: print("Lasso Regression RMSE:", lasso_rmse)
Lasso Regression RMSE: 1.5455915824623936
```

Model Comparison and Selection:

After evaluating each model individually, the next step is to compare their performance to select the best one. The best model is typically the one with the lowest RMSE or the highest cross-validated performance.

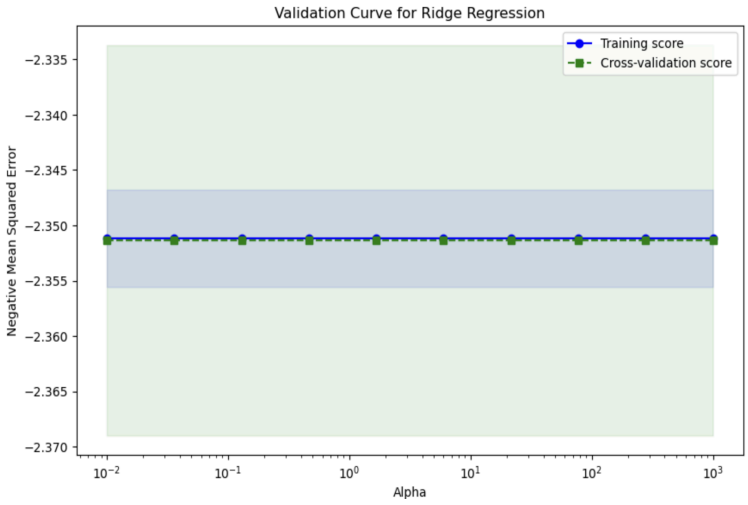


Ridge Regression matches Linear Regression low RMSE and boasts similar cross-validated performance. Its unique ability to tackle multicollinearity and overfitting makes it a robust choice. Thus, Ridge Regression is recommended as the preferred model for its improved stability and generalization.

Model Tuning:

- Tuned hyperparameters of selected models using techniques like GridSearchCV to optimize model performance further.
- Utilized validation curves to visualize the impact of hyperparameter tuning on model performance and identify optimal parameter values.

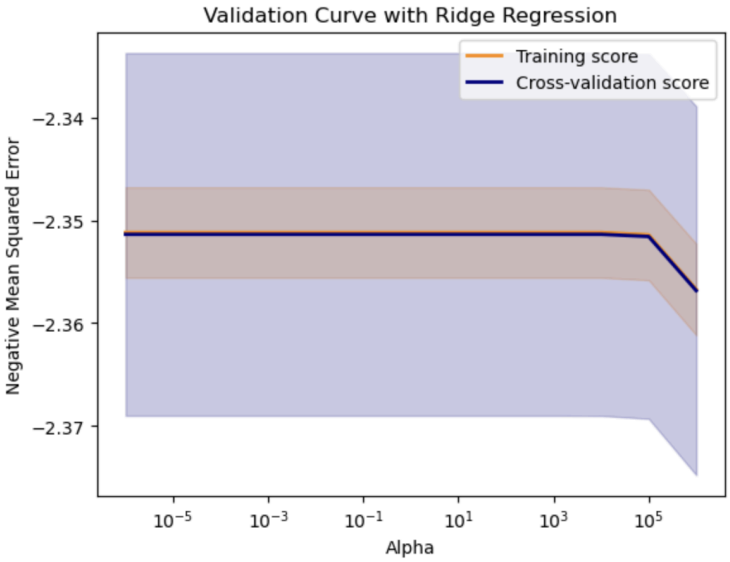
Assessing 1st validation Curve:



Parameter (alpha) range in this case: (-2,3,10)

1. By observing the curve it can be said that the model in this case is overfitting.
2. There is no sensitivity to the hyperparameters.
3. There can also be limitations to the data itself and its impact on the model.

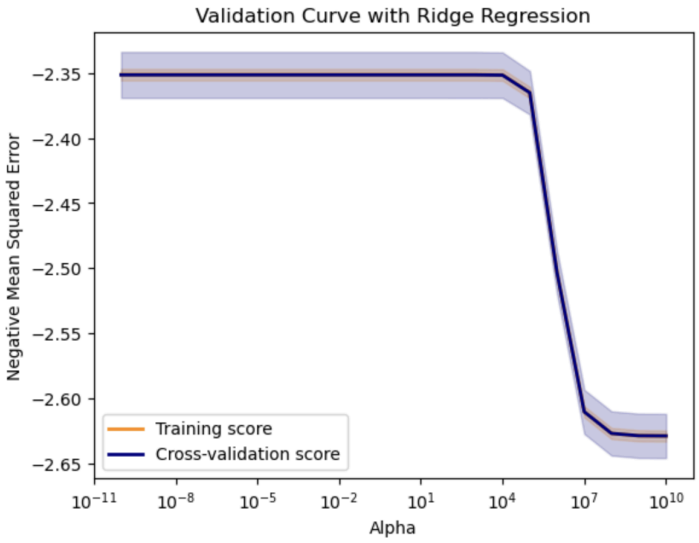
Assessing 2nd Validation Curve:



Parameter (alpha) range in this case: (-6,6,13)

1. Hyperparameter Insensitivity: The model's performance does not significantly change with variations in the hyperparameter within the observed range. This suggests that the chosen range of the hyperparameter may not be affecting the model much. For example, if the horizontal axis represents the regularization parameter (alpha) in a linear model, it might indicate that the model is relatively robust to a wide range of regularization strengths.
2. Saturation Point: The model has reached a saturation point where further increasing or decreasing the hyperparameter does not improve performance. This could happen if the model complexity is sufficient for the data, and adding more complexity (or regularization) doesn't yield better results.

Assessing 3rd Validation Curve:



Parameter (alpha) range in this case: (-10,10,21)

1. The hyperparameter may have a critical threshold value beyond which the model undergoes a significant change. For example, in regularization, a sudden drop might occur when the regularization strength exceeds a point where the model starts to over-regularize, leading to underfitting.
2. The sudden drop could be due to the model encountering a region of the hyperparameter space where it becomes overly sensitive to data sparsity or noise.

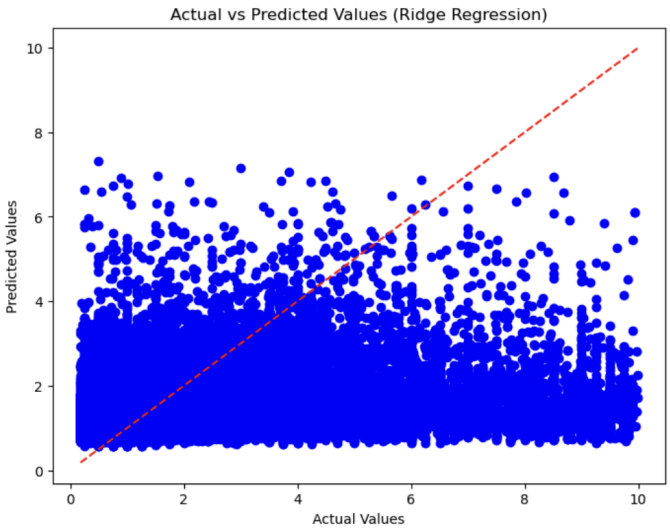
```
Best Alpha: 100.0
Best RMSE: 1.5334127298473492
```

Final Model:

After following the above mentioned methods for scaling , regularization and tuning, the final model with the most effective Hyperparameters and lowest RMSE was selected i.e Ridge regression model with 100 as the optimal alpha value.

This model is further used to predict the job duration of specific farming jobs.
The predicting power of the model is illustrated in the graph below.

R-squared for Ridge Regression: 0.1021914023764583



Prediction power: 10.2%

Learnings:

1. Even though the Ridge regression model was considered the best amongst all other models the model still holds a very low predictability percentage.
2. This can be because the general features itself are not sensitive enough to alter the y value and the correlation between them is not the best.
3. This provides us the insight to have to change some of the features which are more sensitive to the model.

If we have more time, what would we do with the project?

- Explore additional features and interactions that could enhance predictive accuracy.
- Incorporate external data sources (e.g., weather data, soil conditions) to provide more context for job duration predictions.
- Experiment with advanced machine learning models, such as Gradient Boosting Machines (GBM), to potentially improve performance. Additionally, apply regularization techniques to the K-Nearest Neighbors (KNN) method for further refinement.