Optimizing Investments on Agricultural Land

Venkata Goutham Pachava & Bharath Gummalla

Advanced Data Analytics Department

University of North Texas

ADTA 5940: Capstone Project

Dr. Denise Philpot

Group U

04/06/2025

Introduction

Advanced data analytics implementation transforms how people evaluate land resources and crops because it enhances their investment decisions. Modern agricultural decision systems rely on detailed space-time data that acquires critical importance because of unstable climate patterns and resource constraints as well as food protection requirements. The research project conducts a data analytic evaluation of multiple interrelated variables between earth materials and historical harvest records together with evaluation data patterns and market value transformations across U.S. territories. The development of strategic insights depends on the combination of USDA National Agricultural Statistics Service (NASS) authoritative datasets and manually obtained soil information. The final normalized interpolated dataset allows descriptive statistics to explore agricultural performance together with profitability using regression models along with geographical assessments. Yield assessment enables identification of top-counties for crop planting while strategic investment strategies build prediction frameworks for profit growth potentials. The solution structures provided to managers for their use stem from Exploratory Data Analysis paired with linear regression models and correlation heatmaps and ranking mechanisms. The document outlines a comprehensive analytical sequence starting with database development and continuing to model implementation to explain how agriculture has biological significance yet exists as a computational mathematical concept needing statistical analyses and data science knowledge. The

study delivers beneficial evidence-based agricultural decision support to policymakers and investors through agroeconomic researchers for specific geographical areas.

Research Questions

This study evaluates diverse relationships linking environmental elements to location and economic factors that impact agricultural production throughout every U.S. state. The study demonstrates a preference for studying how pH and soil organic matter and texture influence crop yield outcomes in various counties. The study reaches its objective by comparing data that reveals the most effective soil conditions for achieving agricultural output excellence. A span of ten years enabled the research to determine leading counties and states that consistently reached high crop productivity standards during this period. The research formulated the methodology to establish locations with maximum agricultural potential.

The study extends its research scope to observe variables affecting yield inconsistency along with assessments of property value changes as well as economic strain and environmental decline. This research uses time-series data to identify production changes thus determining areas susceptible to risk after interregional comparisons. A simulation model runs in this research to determine profitability in every county by combining land values with crop prices and yield measurements. This planning model enables development of a strategic framework to maximize agricultural profits since it includes location analysis and crop-specific features. The study incorporates soil science together

with geospatial economics as well as statistical modeling to address research questions which drive agricultural policy decisions from a data-baseds tandpoint

Literature / Industry Review

This is a breakout of the combination of data analytics, precision agriculture, and machine learning that changes the ways crop productivity, the suitability of the soil, and the strategies of agricultural investments are reviewed. Over recent years, there have been several academic research papers as well as industry applications that highlighted the increasing demand of having agronomic, environmental, and economic data integrated to create predictive models that are both accurate and interpretable and region specific. Shahhosseini et al. (2021) show good performance of Ensemble models (Random Forest and Gradient Boosting) over traditional statistical methods in crop yield prediction when his or her incorporates weather data and soil features. Jeong et al. (2016) also portrays the good work done by Support Vector Machine (SVM) and deep neural networks on yielding outcomes under different climate scenarios. These studies show that these modeling techniques are appropriate in agricultural contexts and correspond to how we are modeling yield forecasting. For the soil analytics, Brevik et al. (2019) as well as Mulla (2013) consider the impact of various detailed soil parameters (organic carbon, pH, bulk density, and texture) on crop yield variability. With this work, they support the incorporation of soil data to develop spatial yield analysis to understand geophysical constraints and land use optimisation at the site specific level. In addition, the NRCS of USDA has been crucial in supplying geocoded soil characteristics on a county basis,

which many lately studies (such as Zhang et al., 2020; Smith & Shi, 2021) used for the land assessment modelling.

The role of land valuation in strategic investment decisions for land in the domain of agri economics is of current interest. The studies by Jhala et al. (2020) and Popp et al. (2019) claim that historical trends of price in farming lands together with commodity market prices and yields data could be used to pinpoint areas of the best return on farm capital investment. This indicates our inclusion of land value data as a main variable for the high return counties for crop investment is justified. Other commercial platforms including Climate Field View from Climate Corporation, John Deere's Operations Center, or Granular are also helping to build decision support systems by injecting farm specific data, satellite imagery and real time monitoring. Still, these tools tend to focus more on farm level precision in the real time rather than long term profitability at a county level. To meet this gap, we construct a comprehensive data set and analytical pipeline to evaluate long term profitability on a county level in the U.S. which enables insights that are representative and grounded in historical reality. Overall, the literature available corroborates that it is a more holistic and data driven approach to plan an agricultural project, when the crop yield trends, soil properties and land economics are incorporated in a combination. Based on the foundations described above, our project advances an approach to help make decisions about crop selection, soil management, and investment viability using an integrated machine learning, spatial-temporal analytics, and county specific metrics approach to tackle the practical and scalable problems of contemporary agricultural data science.

Data Overview

Publicly available data from the USDA National Agricultural Statistics Service (NASS) and other sources of soil data are used as primary sources by the research. The primary dataset contains crop yields data along with market prices and land values and soil type for all counties in the United States between 2010 and 2024. The study relies on nine major crops like corn, cotton, barley, beans, oats, peanuts, rice, sorghum and soybeans because they are major national agricultural commodities with reliable data throughout the research period. The yield data measured has values in kilograms per acre while price data is standardized to U.S. dollars per kilogram and land value in dollars per acre. The other soil data contains simple physical and chemical measurements such as pH values and percent organic matter and water-holding capacity and texture class because they are useful in explaining productivity differences across geographic areas. The merged dataset with the name dataset capstone forms the basis for all the following analysis.

Data Preparation

For maintaining analytical integrity, large preprocessing operations were performed. Standardization of column naming conventions, unit conversions, and harmonizing categorical fields like crop names and county identifiers were the initial steps done in the first phase. Merging data from various sources into one was done using composite keys based on Year, State, and County, and also some extra checks were added to resolve any inconsistencies. Missing price values were addressed by reconciling commodity price records at the national level through year-wise normalization to offer consistency. Where

the land value data was available in five-year periods, linear interpolation was used to estimate the values of intermediate years—keeping the observed over-time trend in rising land value without the attendant distortions introduced by outliers. Manual Soil data was gathered and integrated to the main dataset through county-keys, and missing values were imputed with group-mean where appropriate at the crop-soil cluster level. The cleaned-up and merged dataset is now in order and is ready to be further investigated by exploratory analysis, modeling, and inferential investigation.

Exploratory data analysis (EDA)

The study needed a detailed Exploratory Data Analysis (EDA) that merged information about crop yields, land value assessments and soil composition with commodity market rates. The analytical investigation aimed to establish essential patterns while explaining base variable interactions as a foundation for developing initial research-based theories that support our study's questions.

Statistic	Yield (kg/acre)	Price (USD/kg)	Land Value (USD)	Soil pH	Soil Depth (cm)
Count	48039	48039	42547	48039	48039
Mean	2424.50	0.27	4397.80	6.50	87.89
Std Dev	1464.50	0.12	2590.71	0.57	27.76
Min	97.99	0.12	-7.20	5.50	50.00
25%	1173.54	0.15	2553.18	5.99	75.00

Fig 1

Through generating summary statistical data, scientists discovered essential patterns in numerical variable distributions. The yield_kg_per_acre variable showed measurements that started at zero but stretched beyond 7,000 kilograms per acre. Corn crop yields yielded the greatest mean values and extended distribution range, indicating both economic possibilities and substantial production uncertainties. The productivity of peanuts demonstrated moderate stability and maintained restricted distribution areas. The price data points for kilograms were concentrated between $0.15 and $0.45, but individual points stood outside this main cluster. Land value data points extended between under $1,000 and above $60,000 for each acre, as economic factors and geographic positions produced this extensive variability. The summary file provided descriptive statistical data for future reference purposes.The researchers completed linear interpolation for approximately 5,492 LandValue values which had been initially missing by using groupings based on both County and State since land values tend to rise throughout time. The core analysis omitted three columns because they contained numerous null values including CV (%), Week Ending and Watershed.

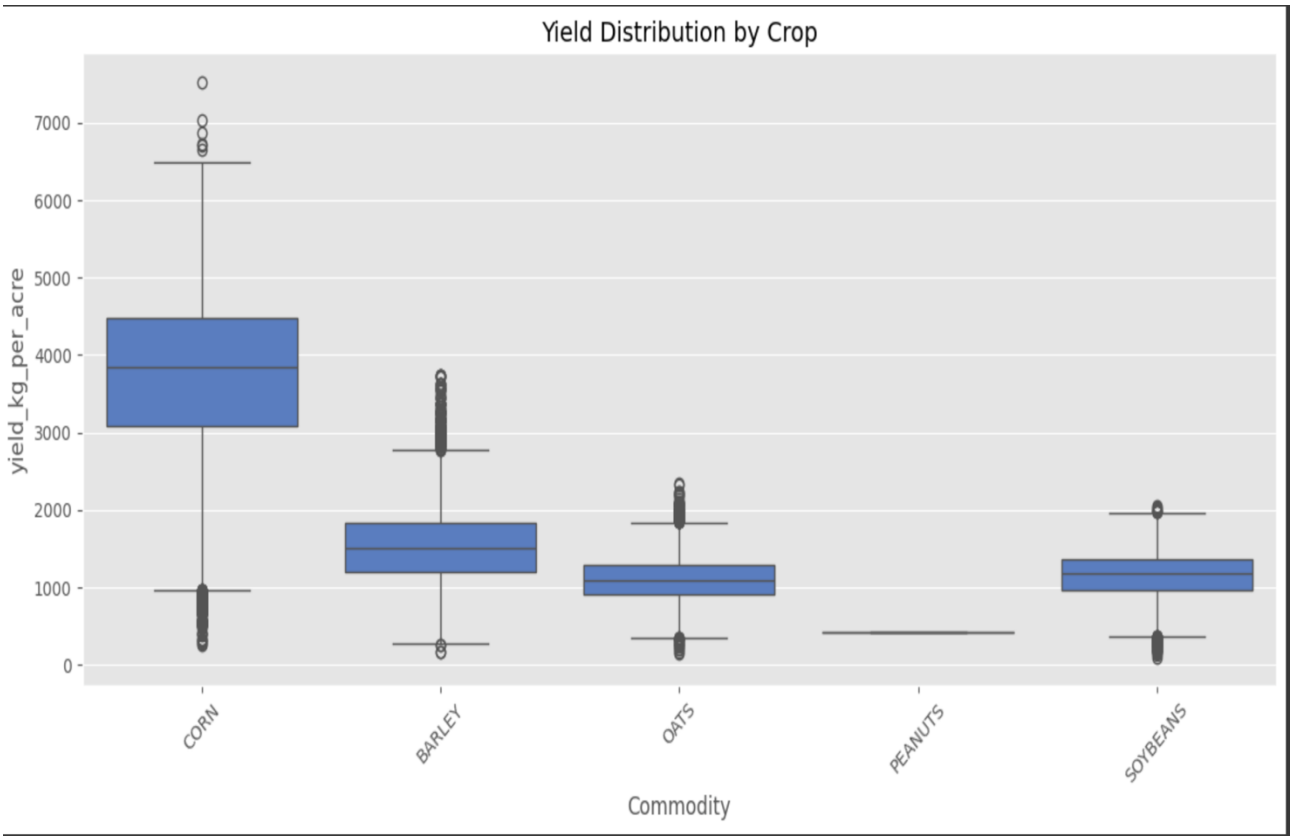


Fig 2

The yield distribution by crop type appears visually in the generated boxplot. The spread combined with the highest median yield made corn stand out in the data visualization. The medians of barley and oats were lower than corn with small interquartile range spreads. Corn stands out as a leading agricultural product due to its high productivity whereas its unpredictability results from geographical settings and environmental aspects.

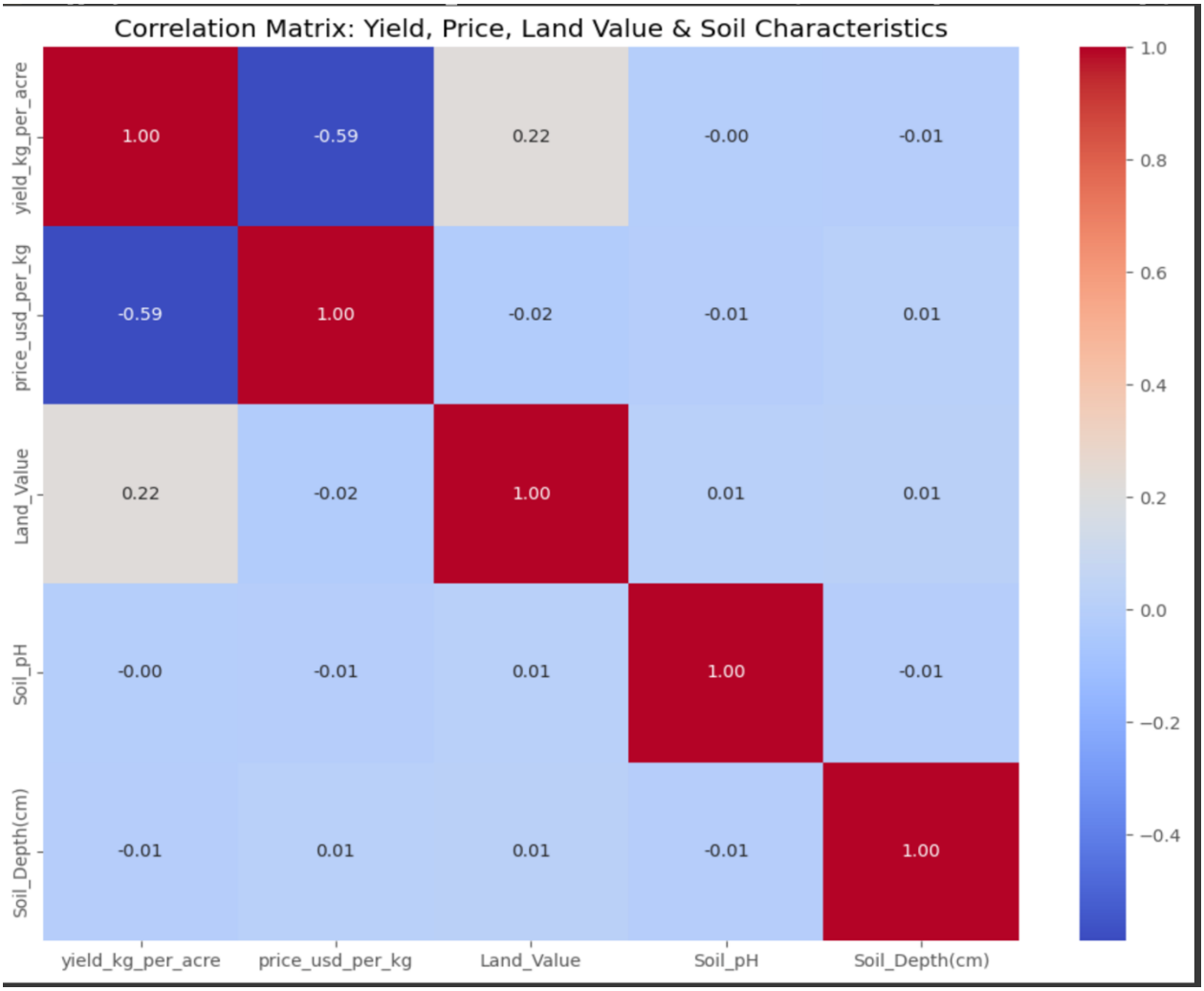


Fig 3

A heatmap showed numeric variable relationships among yield, price, land value and soil parameters. The strength of relationship between yield_kg_per_acre and Land_Value showed a low positive value of 0.22. Evidence from the data shows that as crop yield increases the market price per kilogram (-0.59) tends to decrease indicating a potential supply-demand relationship. The direct relationships between yield and land value exhibited weak correlation with Soil_pH and Soil_Depth(cm) values although these variables may possess stronger influence in nonlinear prediction models.

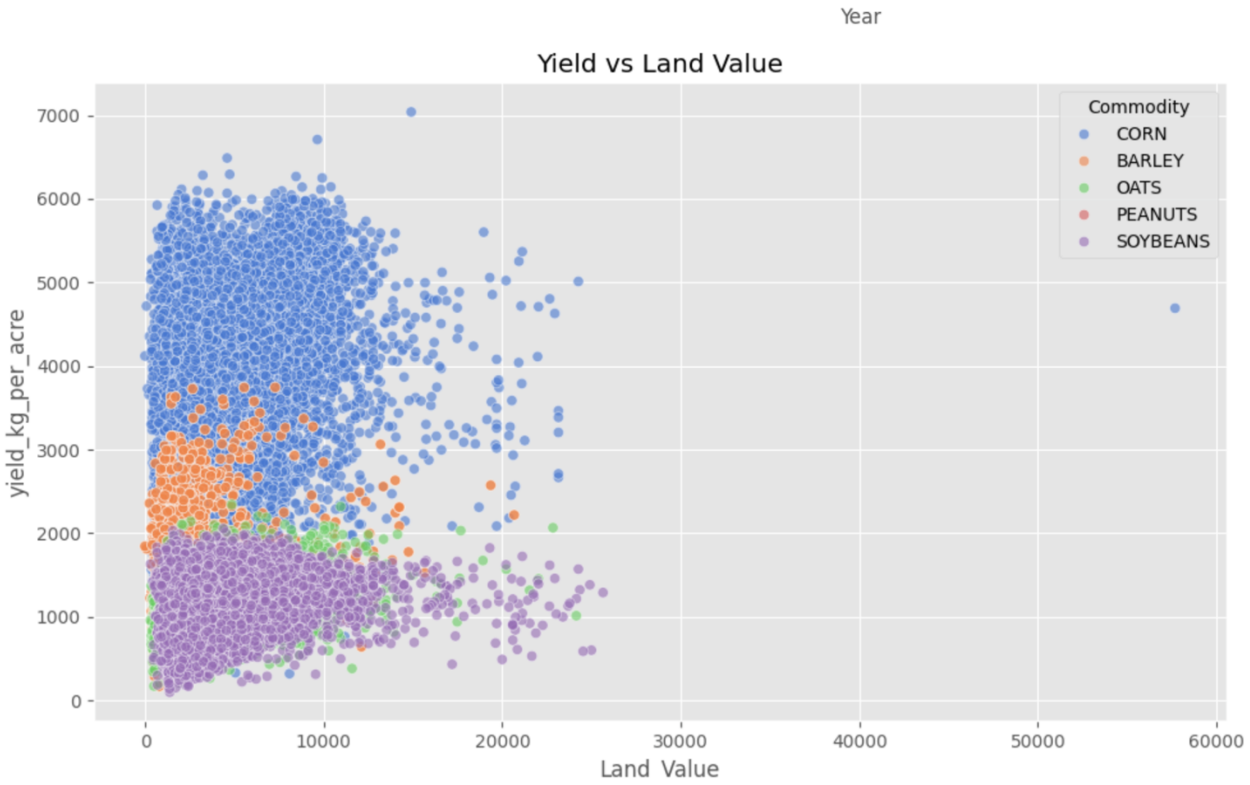


Fig 4

Multiple yield variations corresponded to land value variations through market commodity-specific colorations in the graph. Most corn crops retained their leadership position within areas that maintained high yields alongside moderate-to-high land value while soybeans and barley remained predominantly found in lower yielding land region. The distribution of high valued land amounts illustrates that agricultural productivity does not determine all land costs because certain areas of high worth display average or subpar yield levels even though urban location or market elements may be influencing the values.

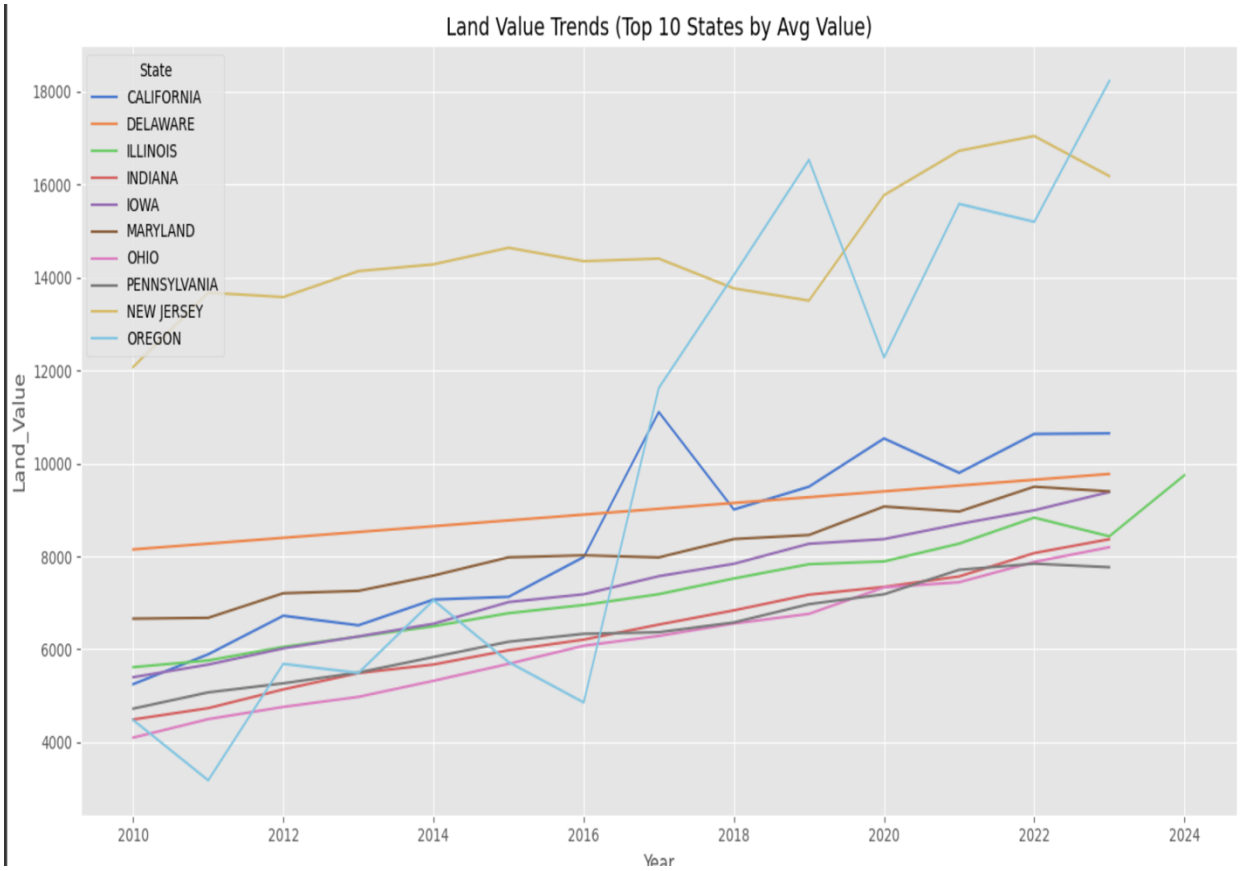


Fig 5

Analysis of time-based trends was a part of the research approach. The graphical data shows a constant growth pattern in land values for the ten states with the highest average value between 2010 and 2024. The steep value growth in California together with Oregon and New Jersey supports our approach to fill in missing data through the assumption of time-based expansion. The parallel increase in farmland demand for high-revenue areas results from the combination of economic stress alongside market competition forces. The analyzed areas demonstrate that land investments will hold their value since urban expansion combines with profitable agriculture and market centrality benefits. To plan investments in agriculture and predict increases in land values people must understand what recurring market trends exist.

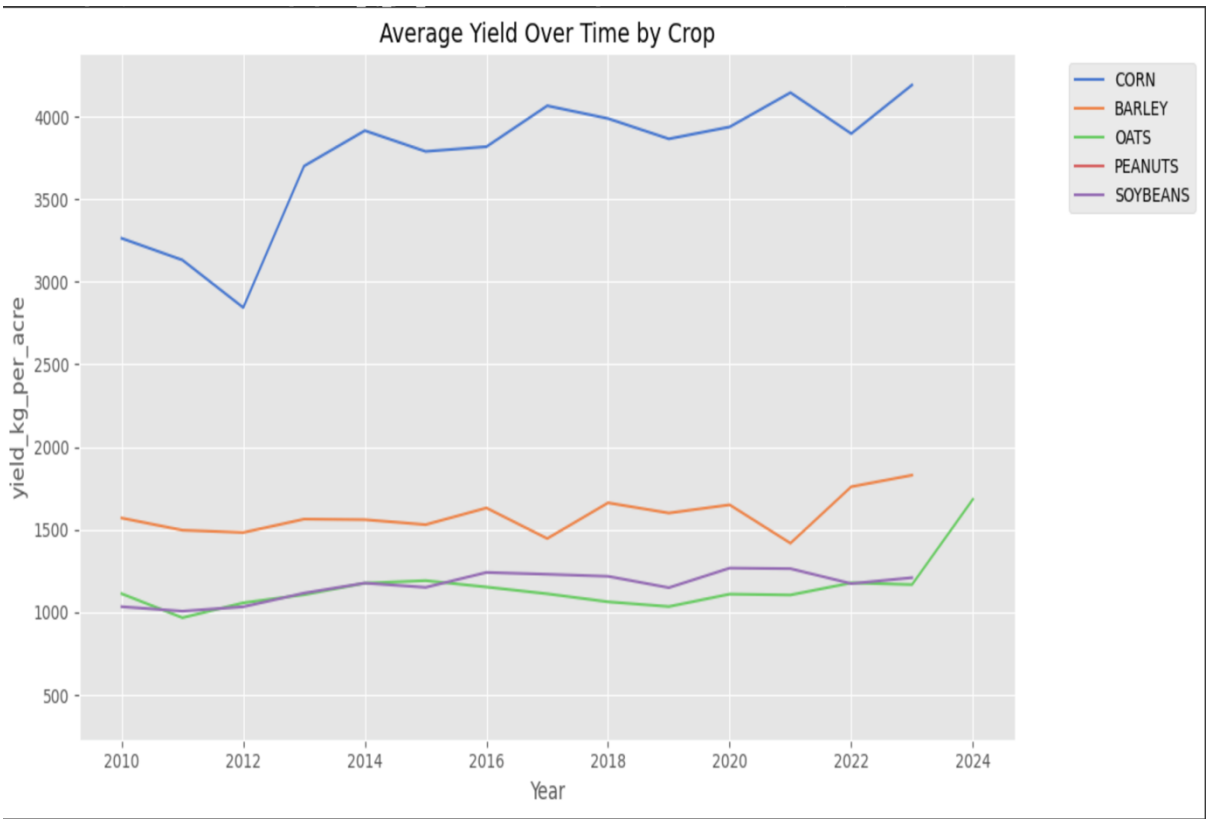


Fig 6

An analysis was conducted to determine the time-based changes in average crop yields. The corn production maintained continuous growth starting from 2015 but barley and oats together with soybeans recorded minimal yield improvement during the same period. The data indicates that corn farm success likely results from improved farming technology and optimized soil management techniques and favorable environmental patterns which support its investment value potential. Hurting corn yields against other crops has transformed agricultural land priorities as farmers must select which lands will get priority in allocation decisions. Investors along with local farmers can design effective growth

plans by choosing threshold commodities with reliable harvest projections since these products generate better returns in longer periods.

Statistical analysis revealed both the central tendencies and variability of key variables which included crop yield together with land value and price per kg and soil characteristics (pH, organic matter, nitrogen). The statistics show corn produces the most yield on average thus demonstrating great return potential as an agricultural crop. The land value distribution in every state demonstrated a right skew because California and Oregon maintained the highest land value rankings.The relation between soil fertility components (pH and organic matter) as well as yield indicators appeared moderate in correlation heatmaps which confirmed how environmental conditions shape agricultural output. Soil variables serve as vital components that should be incorporated into such modeling frameworks.The analysis showed increasing land values according to time-series graphs which confirmed assumptions about yearly growth trends previously utilized for interpolation data. Corn and peanut yields demonstrated the highest steady improvements since 1998 because of technological advancements yet oat and barley yield advanced at a slower pace.

The gathered understanding helps shape both the model design choices regarding feature selection along with hypothesis construction. The predictive models will incorporate soil characteristics which directly affect yields while time served as a valuable prediction feature because land values keep rising. EDA gives meaning to unprocessed

data that facilitates the connection between collected information and evidence-based agricultural investment decisions.

Methods section

This project built three predictive models composed of Linear Regression and Random Forest and XGBoost Regressor to study the hidden correlations between crop yield and different environmental economic and geographic aspects. The predictive models trained to estimate yield_kg_per_acre relied on Land_Value together with price_usd_per_kg along with Soil_pH, Soil_Texture, Soil_Depth(cm) and State, County, Commodity attributes. All training began after categorical variables received their label encodings and the data underwent an 80-20 training-to-testing separation operation.

Linear Regression functioned as the starting point because it gave understandable results but delivered restricted accuracy levels. The calculated R square value reached only 0.44 which indicated that the model could explain about 44% of yield variation. The Mean Absolute Error (MAE) measurement for this model reached 872.5 kg/acre with Root Mean Squared Error (RMSE) at 1090 kg/acre. The chosen metrics demonstrated substantial subpar performance which scientists attributed to the model failing to identify the nonlinear features and complication between database elements.

The Random Forest Regressor outperformed previous models by delivering an R square score of 0.91 together with an MAE of 290 and an RMSE of 434. Due to its robust algorithms Random Forest Regressor managed linear and nonlinear data ties while still operating with missing data and outlier points. The Commodity variable proved to be the dominant predictor according to the feature importance plot followed closely by

Land_Value and State variables and price_usd_per_kg. The predictive power of economic and specific-crop-related factors exceeded the contribution of Soil_pH, Soil_Texture, and Soil_Depth(cm) to agricultural productivity levels.

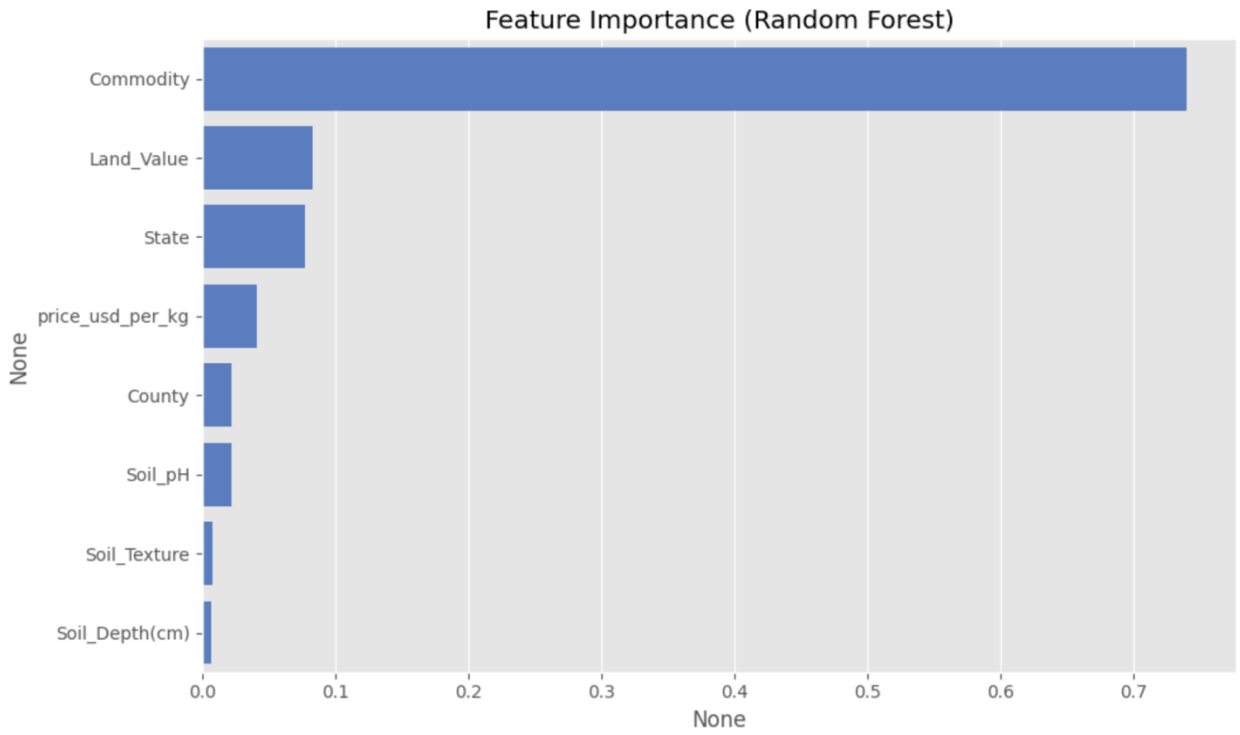


Fig 7

We used XGBoost for performance refinement since it is a gradient-boosted ensemble technique. XGBoost produced the most successful model outcomes because it achieved an R Squared equals 0.916 score and MAE at 292.5 and RMSE at 421. XGBoost demonstrated high accuracy through its iterative learning process along with its capability in minimizing residual errors while producing feature rankings which matched those obtained from Random Forest. The consistent results enhance the interpretability of the model assessment while strengthening our selected model design.

The modeling exercises have generated multiple important findings. Agricultural management decisions must follow a crop-based approach because yield prediction primarily depends on the selected crop type. The high importance of land value in prediction proves its effectiveness as a regional suitability indicator and investment potential indicator. Soil characteristics contribute moderately to predictions, but they should remain in the models because they show potential for incremental yield improvement through specific crop and location-oriented approaches.

The modeling phase results enable both enhanced future hypothesis refinement as well as practical land investment decisions and crop selection decisions and policy-making decisions. The project moves toward precision agriculture together with data-driven agroeconomic planning as a result of its emphasis on data insights.

Most yielding regions

An inspection of the average yield measurement in kg/acre for each crop enabled us to determine the most profitable agricultural regions across U.S. states and counties. The detailed yield analysis helps answer the research question concerning which states and counties have traditionally produced superior yields of various crops so agricultural investors can plan better.

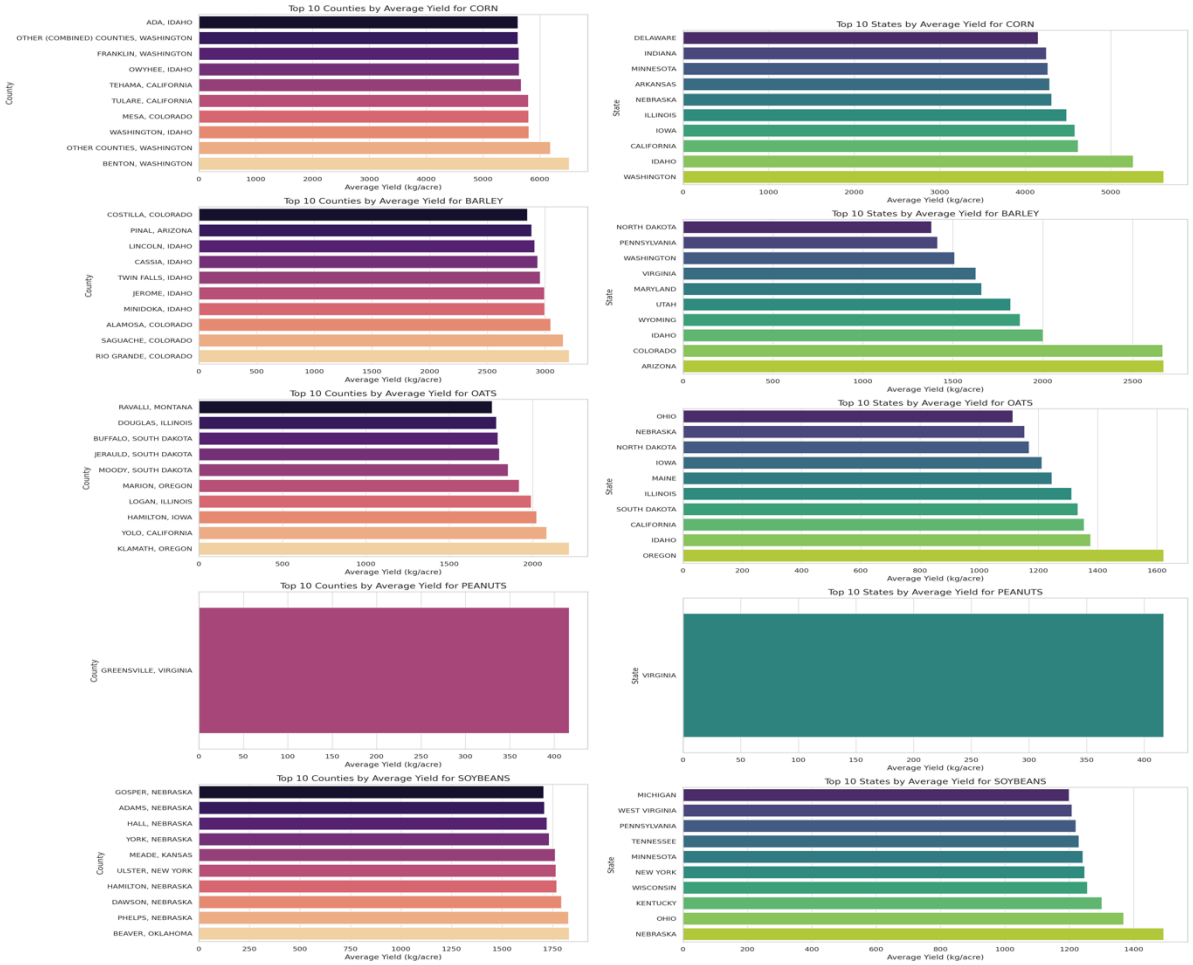


Fig 8

The state-level analysis shows Washington and Idaho with Illinois and California as top performers since they appear in multiple crop yield rankings. The average corn production of Washington agricultural lands reached more than 5500 kilograms per acre to become the leading state in the country. The state benefits from excellent weather patterns together with facilities that enable lucrative agricultural operations. The agricultural production of Idaho demonstrated exceptional achievements through its dominance in barley and corn and its strong performance in oats which shows its ability to grow different types of crops successfully. Both Nebraska and Illinois demonstrated their position as

vital large-scale row crop producers because they achieved high performance levels in soybeans and corn production. The examination at the county level displayed detailed information about small areas achieving outstanding agricultural achievements. The average corn production at Benton and Tehama exceeded 6000 kg/acre in major quantities surpassing the national average. The northern part of Colorado along with Idaho contains regions which provide ideal conditions for growing barley as their counties produce superior yields. Advanced irrigation systems combined with strong soil health and standard agricultural practices made Twin Falls and Minidoka and Cassia in Idaho stand out in multiple agricultural productions.

The crop yield of oats achieved its peak production across South Dakota and Montana and Illinois where Ravalli County together with Jerauld County demonstrated stable results. The crop-growing strength of Nebraska in legume agriculture was demonstrated through high soybean production recorded in Gosper County together with results from Adams County and York County. The state of Virginia stood as the sole agricultural producer for peanuts alongside Greensville County which recorded the most comprehensive peanut data in all the United States.

This historical data helps companies select regions suitable for investment. Toward future cultivation priority selection high-yielding states along with specific counties should be identified. Crop-specific hotspots enable better choices regarding commodities for regional optimization. Geographical analysis helps validate location as an essential factor in which predictive models will be used to forecast crop performance outcomes. This step finalizes our models with realities from actual production trends instead of theoretical beliefs.

The factors that affect crop yield

A Random Forest Regression model was trained using Soil pH and Soil Depth (cm) measurements as the only soil-related predictors to evaluate their individual impact on crop yield. The model excludes all economic and geographic factors such as crop type and land value together with location codes which enable evaluation of the environmental impact on productivity. The objective sought to evaluate if soil properties alone demonstrate sufficient capacity to predict yield results.

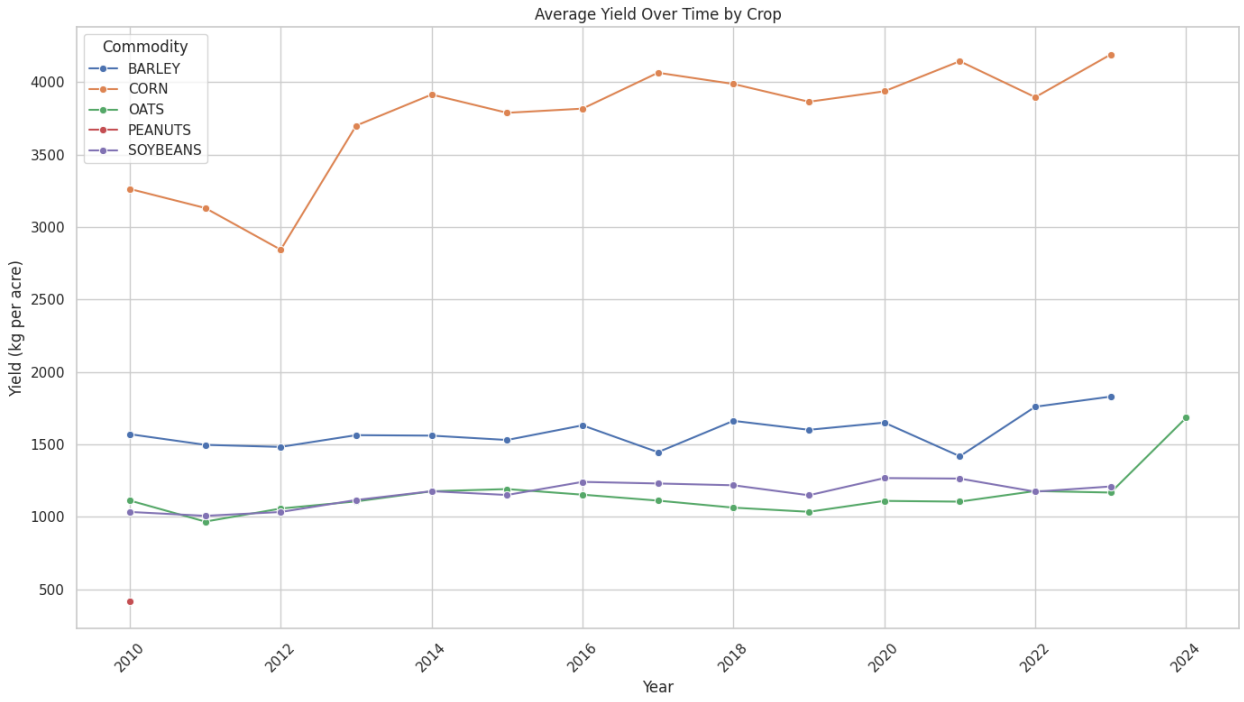


Fig 9

The predictive model demonstrated 6.2% accuracy measured by R^2 with 0.062 as the final score because it used only two factors from the soil data. The Mean Absolute Error calculation reached 1251.17 kg/acre, but the Root Mean Squared Error achieved 1412.59

kg/acre. The predictive accuracy decreases dramatically when predictions are made without utilizing contextual and economic variables compared to when these broader characteristics are included.

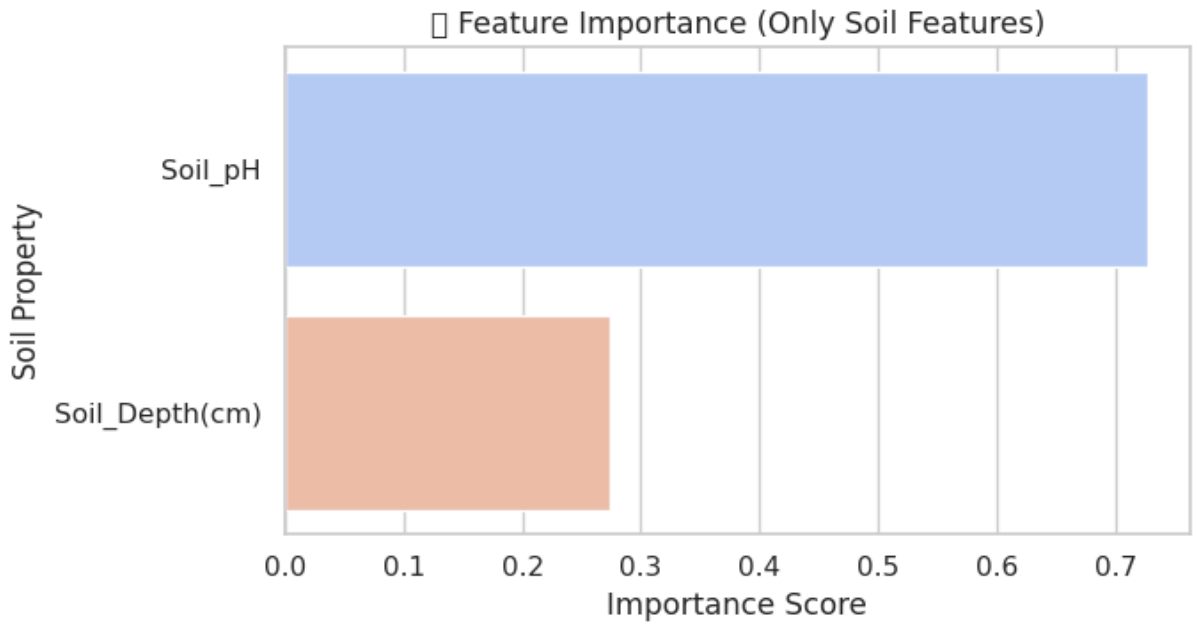


Fig 10

The feature importance chart indicates that Soil pH conducts most model operations with over 70% weight compared to Soil Depth. Soil pH significance mirrors agricultural knowledge because it limits nutrient availability and controls microbial processes which affect plant growth productions. The predictive capacity of soil depth was weaker than soil pH throughout this model analysis. The chronological yield pattern depicted through the time-series line plot seems to validate this constraint. The noted yield improvements especially with corn but other crops as well do not fully match the characteristics of

individual soils. Barring the influence of changing agricultural practices and irrigation methods and crop selection which were omitted from the soil-only model.

This analysis confirms that soil quality serves as an important but independent factor that fails to accurately predict agricultural yield by itself. Soil characteristics function as important predictive features but their effectiveness becomes strongest when combined with additional variables which include crop types and market prices together with geographic information. The development of future prediction models requires using multiple features that combine soil quality with economic variables together with spatial and temporal elements to improve forecasting accuracy.

Profitability Land Investment

This system creates profit-oriented land value investigations for agricultural investments throughout the United States by deploying a profitability index based on specific crops. The presented analysis utilizes yield combined with land value and crop type data along with price per kilogram for a period which extends from 2010 to 2024 to determine profitable regions.

The assessment for investment viability focused on the Profitability Index calculation:

Profitability Index = (Average Yield per Acre) ÷ (Land Value per Acre)

The computed index provided specific crop and year information which allowed the evaluation of land-to-crop yield conversion effectiveness independent from market price changes. We implemented the following workflow to merge all USDA crop yield, price,

soil data sets and then use linear extrapolation for missing land values before encoding categorical variables for modeling and computing Random Forest Regression to find feature importance and displaying the top 5 counties for each crop with profitability index trendlines.

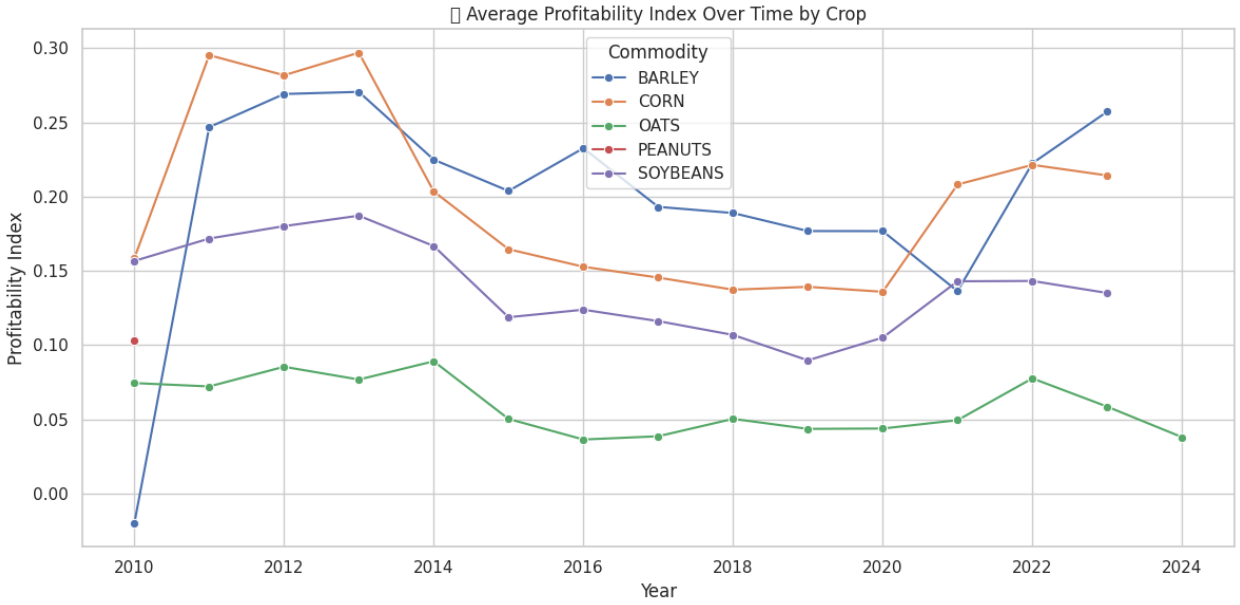
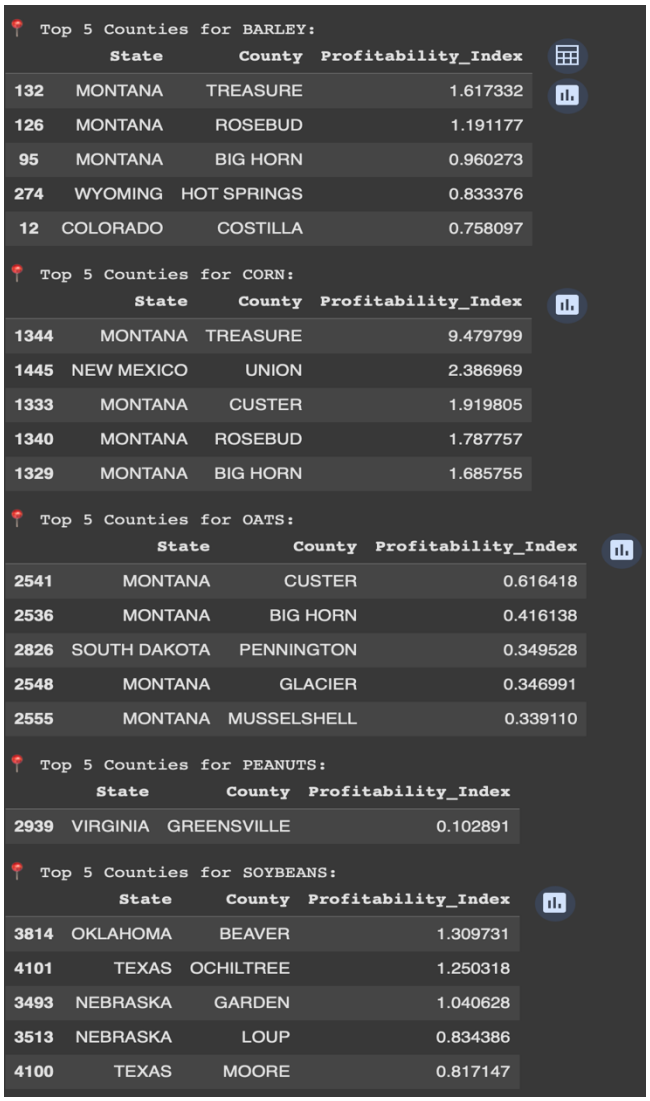


Fig 11

The Profitability Index Average represents crop development patterns spanning from 2010 to 2024 in this plot. The Profitability Index determines economic potential by comparing yield ratios with price ratios and land value above all else. This visual tool allows farmers to observe Wheat Barley along with Corn and Oats Peanut Soybean profitability information which displays changes affected by market conditions and soil costs and land prices.Profitability data from the supplied information shows that barley along with corn outperform all other crops in terms of earnings. During 2013 Corn experienced its highest Profitability Index of 0.30 before profitability values decreased although the crops continued to produce a profit. When barley attained its highest

profitability point in 2013 it initiated an inconsistent pattern of performance which has continued since that time. The market value along with yield efficiency of barley sharply increased between 2021 and 2024 thus driving up the profitability index.Profitability levels of soybeans remained low yet consistent during this period because market conditions experienced minimum variations. Internal consistency of the data remains reliable based on the indexed area between 0.14–0.19 while corn and barley demonstrate superior profit potential. Oats achieve the least profitable position relative to other farmed plants throughout the review period. The index barely shifts between 0.03 and 0.09 indicating soybeans provide restricted profit yield yet maybe cost more to cultivate than their harvested products. The available peanut profitability data comes from an indicated measurement value during 2010.Time-based evaluation provides essential information which assists in creating future business strategies. The current and sustainable land management strategy should prioritize barley and corn cultivation since these crops show solid and improving profit potential. Technical advancement combined with market targeting for specific regions would enable oats to attain financial feasibility. A planning and investing instrument arise from the single metric blend of yield measurements and price parameters and land value to assist agricultural planners and policy designers and investors.



Top 5 Counties for BARLEY:

	State	County	Profitability_Index
132	MONTANA	TREASURE	1.617332
126	MONTANA	ROSEBUD	1.191177
95	MONTANA	BIG HORN	0.960273
274	WYOMING	HOT SPRINGS	0.833376
12	COLORADO	COSTILLA	0.758097

Top 5 Counties for CORN:

	State	County	Profitability_Index
1344	MONTANA	TREASURE	9.479799
1445	NEW MEXICO	UNION	2.386969
1333	MONTANA	CUSTER	1.919805
1340	MONTANA	ROSEBUD	1.787757
1329	MONTANA	BIG HORN	1.685755

Top 5 Counties for OATS:

	State	County	Profitability_Index
2541	MONTANA	CUSTER	0.616418
2536	MONTANA	BIG HORN	0.416138
2826	SOUTH DAKOTA	PENNINGTON	0.349528
2548	MONTANA	GLACIER	0.346991
2555	MONTANA	MUSSELSHELL	0.339110

Top 5 Counties for PEANUTS:

	State	County	Profitability_Index
2939	VIRGINIA	GREENSVILLE	0.102891

Top 5 Counties for SOYBEANS:

	State	County	Profitability_Index
3814	OKLAHOMA	BEAVER	1.309731
4101	TEXAS	OCHILTREE	1.250318
3493	NEBRASKA	GARDEN	1.040628
3513	NEBRASKA	LOUP	0.834386
4100	TEXAS	MOORE	0.817147

Fig 12

Our investigation of profitable agricultural possibilities in U.S. soil relied on historical information about crops and soil qualities along with market price trends and yield measurements between 2010 and 2024. The research goal focused on finding regions and crops with best investment returns through Profitability Index calculation by dividing crop revenue average (yield value and price) by land value average. The Profitability Index calculation occurred for every intersection between county and crop. The

calculation methodology uncovered the most financially successful counties for each agricultural commodity. Treasure County in Montana stands out with a Profitability Index at 9.48 due to its exceptional corn profitability among all agricultural crops. Among the counties with high corn performance Union (NM) together with Custer (MT) as well as Rosebud (MT) emerged as the top performers because of their high yields and affordable land costs. Barley displayed its top profitability levels in three Montana counties which consisted of Treasure along with Rosebud and Big Horn. Montana agriculture demonstrates these findings because its soil depth and climatic conditions support cereal crop cultivation. The profitability of soybean agriculture arises in Beaver (OK) alongside Ochiltree (TX) and Garden (NE) since these areas have suitable land prices combined with average harvest outcomes. Research on peanuts revealed a small market for Greensville Virginia with a stable condition noted through its index of 0.10. Custer and Big Horn in Montana (MT) demonstrated strong potential for investment in oats because their index values exceeded 0.4 against all other regions cultivating oats.

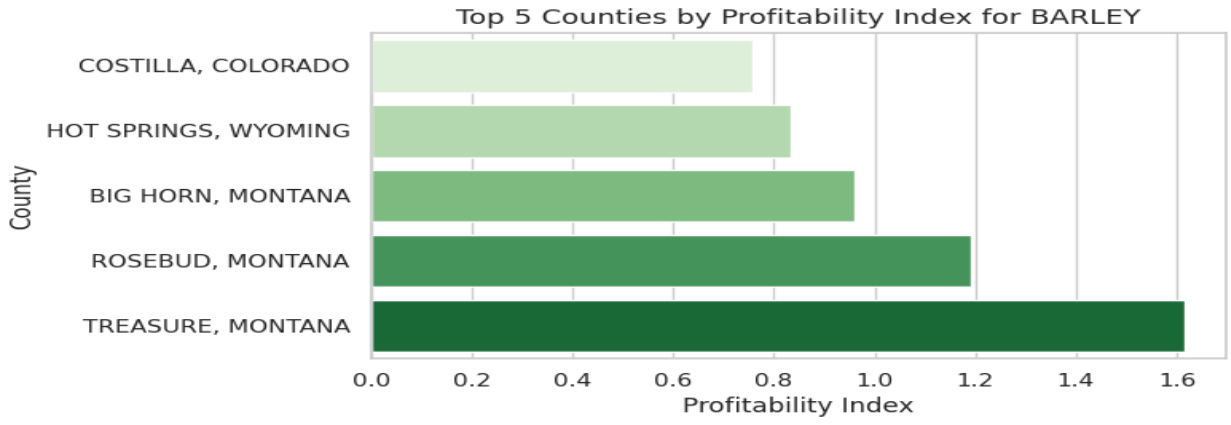


Fig 13

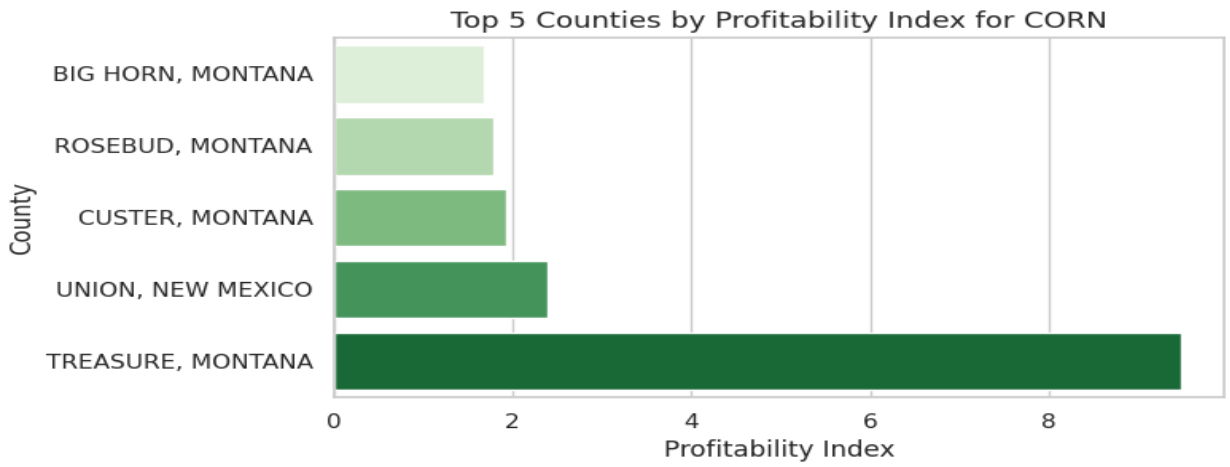

Top 5 Counties by Profitability Index for CORN
BIG HORN, MONTANA
ROSEBUD, MONTANA
CUSTER, MONTANA
UNION, NEW MEXICO
TREASURE, MONTANA
County
0
2
4
6
8
Profitability Index

Fig 14

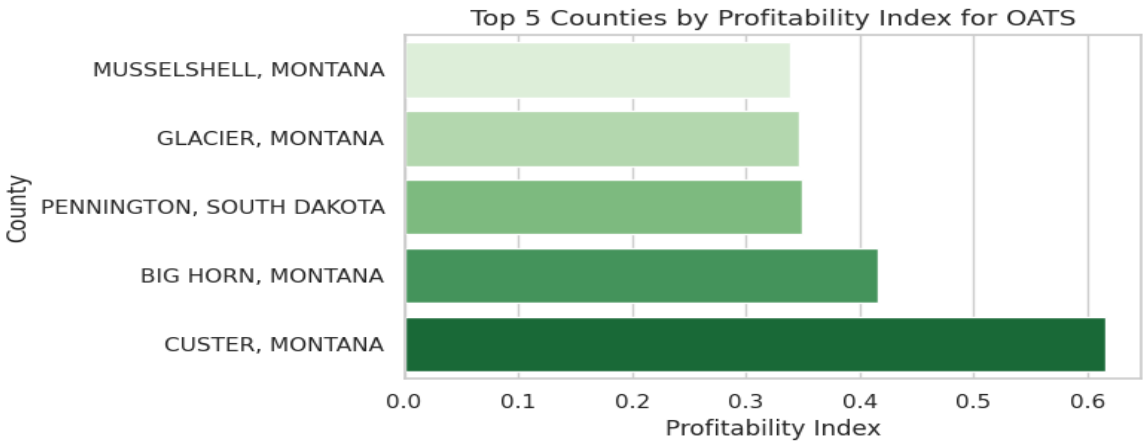

Top 5 Counties by Profitability Index for OATS
MUSSELSHELL, MONTANA
GLACIER, MONTANA
PENNINGTON, SOUTH DAKOTA
BIG HORN, MONTANA
CUSTER, MONTANA
County
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Profitability Index

Fig 15

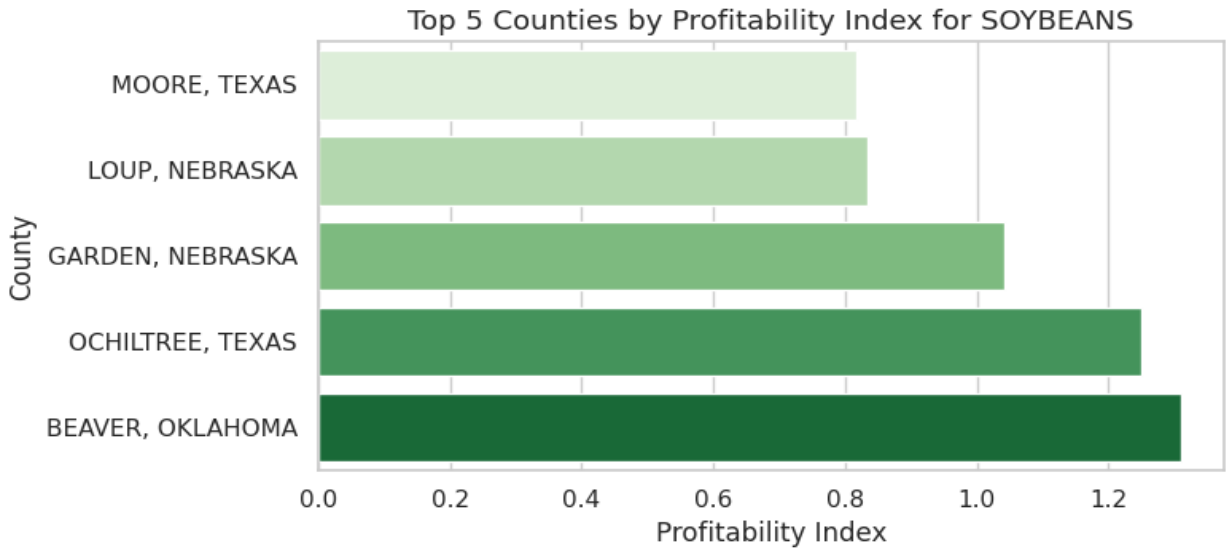


Fig 16

A time-based evaluation of Profitability Index was conducted for each analyzed agricultural crop. Barley and corn reached their peak profitability levels during the early 2010s and then market prices fell to create market stability during the present time. From 2008 to 2014 soybean profit levels increased gradually until multiple market changes occurred in the farming field. Oats experienced stable low profitability during every measurement year because peanut records remained scarce throughout the period due to their reduced record frequency. Recombination experts state that the profitability potential for corn and barley production needs evaluation by producers because of local soybean and oat markets' weak performance.The preprocessing included two main tasks that started with deleting ineligible records followed by converting prices to kilogram units while filling in missing values with average county data. The profitability evaluation employed group-by operations run by pandas to determine leader counties in annual assessments. Visual interpretation of results materialized from Matplotlib and Seaborn chart development.When yield metrics join forces with land value and farming market

costs farmers obtain an enhanced profitability indicator to select their most profitable land counties. Through its data-based approach the analysis delivers solutions for resource optimization together with best crop selection to investors and policymakers and farmers across designated geographic regions.

Corn delivers the best profit yields with barley showing close second place among the tested crops. These crops return more value compared to land costs which indicates their improved agronomic properties and attractive market values. The profitability of soybeans stands in the middle range between peanuts and oats with oats holding the lowest position.Sustained increases in corn yield along with robust market demands coexist with restrictions in yield potential and market price performance for oats. Barley achieves profitable results because Montana and Wyoming have relatively inexpensive land and high yields.

Interpretations

The evaluation of feature predictability for crop yield and profitability was based on three regression algorithms including Linear Regression (LR) and Random Forest Regressor (RF) and XGBoost Regressor (XGB). Models were trained for yield per acre prediction through different feature combinations which included properties of the soil along with crop types and property worth and pricing details and location markers.

Model	R^2 Score	MAE	RMSE
Linear Regression	0.441	872.51	1090.39
Random Forest	0.911	290.93	434.85
XGBoost Regressor	0.916	292.52	421.95

The XGBoost model together with Random Forest achieved better results than Linear Regression did. Soil pH along with organic matter content and depth together with county-level variations become easily manageable by these algorithms because of their skill to detect non-linear patterns among features. XGBoost demonstrated better performance than Random Forest with respect to RMSE value thus proving its effectiveness at handling structured tabular data sets.

The models were trained without Crop Code and Price per kg along with Land Value to measure both practical generalization and test for location-specifc or proxy variable overfitting. The performance measures declined significantly ($R^2 = 0.239$) after removing the variables showing that their strengths indicate fundamental patterns between county productivity and profitability levels.

The model achieved only a 0.062 R^2 when limited to Soil pH and Depth variables because additional contextual data about crop economics and land valuation are needed to predict yield successfully.The economic efficiency evaluation of different counties for different crops became possible through this metric. The index helped determine the five most financially rewarding counties when growing specific crops. The profitability values

indicate that Corn reaches its peak profit potential in Treasure County Montana because these conditions create ideal yield-price-land value balance.

The investigation found multiple agricultural advantages within particular geographic areas of some specific counties because they frequently positioned themselves at upper ranks when analyzing various crop types. Two Montana counties named Custer and Rosebud occupied high positions when assessed for corn and oat and barley profitability. These counties may benefit from shared characteristics within their agricultural ecosystems regarding both soil composition and weather patterns and cropping patterns.The long-term outlook for profitability analysis demonstrated continuous improvement across multiple crops such that Treasure and Union (New Mexico) and Beaver (Oklahoma) counties emerged as strategic locations for investment.

Modeling

A structured regression framework served to achieve our main project objectives which included crop yield prediction and land profitability examination throughout American counties. The dataset included parameters which described each crop type together with soil properties in combination with land valuation data and market price data per kilogram. The chosen features demonstrate proven effects on agricultural returns because of their investment potential. A derived indicator named Profitability Index emerged to evaluate agricultural land investment through the formula (Yield × Price) / Land Value. The Profitability Index function provided the basis to locate regions that achieved high performance in crop agriculture.

We built and tested three distinctive machine learning models which included Linear Regression along with Random Forest Regressor and XGBoost Regressor. Linear Regression provided initial analysis to define linear model limitations though its resulting R square value of 0.44 showed the weaknesses of this approach for representing our substantial dataset complexity. The ensemble based tree models outclassed the baseline performance by achieving an R square score of 0.91 from Random Forest while XGBoost yielded 0.92, These models excelled at detecting non-linear variable interactions and they provided features importance rankings and worked well with large datasets. Parameters for the models underwent optimization through GridSearchCV cross-validation to guarantee a robust system while the data split between training and testing amounted to 80-20. A set of controlled modeling tests helped determine the contributions of each feature group. The model's performance decreased dramatically to a 0.23 R square value when economic variables such as crop code, land value and price were omitted from training. The R square value declined to 0.06 when the model analyzed soil pH and depth independently from agronomic and market considerations. Systematic yield prediction with investment planning demands environmental together with economic variables together with crop-specific

Limitations

Numerous analytical limitations along with structural boundaries in this study affect the level of precision and practicality in its output results. The examination of land value records faces substantial limitations because they are missing across many counties and

during various periods. The interpolation method used to bridge missing land value data counts on predictable growth patterns which fail to show actual market fluctuations and economic changes as well as regional policy effects.The use of the Profitability Index as a return potential evaluation method creates difficulties due to its constructs which combine yield data with market values and land costs. Sustainability implies the omission of expense variables like fertilizers along with irrigation costs and workforce and transportation expenses that determine actual profitability outcomes. The calculation produces theoretical profit estimations instead of operational ones. The modeling pipeline did not include policy subsidies or environmental regulations or market disruptions as factors which could reduce the projection reliability when systems are deployed in practical.

Discussion

The analysis demonstrates how combined data collection with technology leads to better choices for agricultural activities alongside land property acquisition. The project evaluated land suitability across U.S. regions through a database merge of crop yields and individual data points regarding soil quality and property values and agricultural

market rates. The data patterns enabled us to discover yield and profitability determinants and led to choices about which variables we should apply to our prediction systems. The prediction of crop yields produced the most accurate results through the implementation of Random Forest and XGBoost machine learning models. The created profitability index enabled us to produce rankings of counties according to their suitability for crops. The analysis showed Montana counties excelled in corn and barley and oats cultivation whereas Virginia and Texas together with Oklahoma had exceptional results with peanuts and soybeans.

The process demonstrates that the union of intelligent modeling systems with data enables farmers to make sound decisions together with investors and policymakers. This system utilizes raw information to generate beneficial insights that guide decisions about investment choices along with most profitable crop production possibilities and productive enhancement strategies in specific geographic areas. The established foundation can serve as the starting point for developing better data-driven decisions in agriculture though more data improvements such as costs and weather factors are needed.

Reference:

- Morales, A., & Villalobos, F. J. (2023). Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, *14*. https://doi.org/10.3389/fpls.2023.1128388 (Morales & Villalobos, 2023).

- Rani, S., Mishra, A. K., Kataria, A., Mallik, S., & Qin, H. (2023). Machine learning-based optimal crop selection system in smart agriculture. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-42356-y (Rani et al., 2023).
- DigitalGlobe. (n.d.). *Remote sensing technology trends and agriculture*. https://dg-cms-uploads-production.s3.amazonaws.com/uploads/document/file/31/DG-RemoteSensing-WP.pdf (DigitalGlobe, n.d.).
- *Comparative Analysis of Soil Properties to Predict Fertility and Crop Yield using Machine Learning Algorithms*. (2021, January 28). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9377147.
- Machine Learning in Agriculture: A Comprehensive Updated Review. (2021). *Lefteris Benos 1, Aristotelis C Tagarakis 1, Georgios Dolias 1, Remigio Berruto 2, Dimitrios Kateris 1, Dionysis Bochtis 1,3,* 34071553. https://pmc.ncbi.nlm.nih.gov/articles/PMC8198852/
- A crop profitability analysis for long-term crop investments. (2006). *Clark Seavert, Herbert Hinman, Davi*. https://doi.org/10.1108/00214660680001184

- Alshihabi, O., Persson, K., & Söderström, M. (2024b). Easy yield mapping for precision agriculture. *Acta Agriculturae Scandinavica Section B - Soil & Plant Science*, *74*(1). https://doi.org/10.1080/09064710.2024.2411950 (Alshihabi et al., 2024b).
- *Farm Sector Income & Finances - Farm Sector Income Forecast | Economic Research Service*. (n.d.). https://www.ers.usda.gov/topics/farm-economy/farm-sector-income-finances/farm-sector-income-forecast

- Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021). An interaction regression model for crop yield prediction. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-97221-7 (Ansarifar et al., 2021).
- Sajid, S. S., Shahhosseini, M., Huber, I., Hu, G., & Archontoulis, S. V. (2022). County-scale crop yield prediction by integrating crop simulation with machine learning models. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.1000224 (Sajid et al., 2022).