# 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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

## 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 soil information. The research data got from public repositories, which contain crop yields, along with land valuations and soil characteristics information at the county level (USDA Quick Stats -https://quickstats.nass.usda.gov). 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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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. This research identifies most profit land-crop combinations. They are few limitations like price of crops various from place to place that may affect calculation of profitability. Additional analysis about the study's limitations takes place in a later section.

## 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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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-based standpoint

## 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.

Modern agricultural technology implements novel solutions for crop cultivation, and it also changes the process of property evaluation and operational optimization systems. During past times farming choices regarding land and planting relied on both human experience and natural seasonal indicators. Nonpartisan knowledge no longer serves as an effective solution since climate patterns remain unpredictable at the same time markets continue to

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

adapt. Modern research uses combined datasets of agricultural farming factors together with environmental elements and financial components to bring about better long-term decisions through predictive models at local levels.

Crop yield prediction reached a breakthrough by using machine learning model applications. Morales & Villalobos (2023) demonstrate how ML-based biophysical crop simulation systems use environmental linkage to farming elements for improving predictive accuracy. A hybrid machine learning model proposed by Rani et al. (2023) optimized crop selection decisions through integration of soil and environmental as well as economic elements to boost yield production estimates. The forecasting systems deliver superior results compared to conventional approaches thus enabling stakeholders to use data-driven strategies to make decisions about agricultural investments while improving financial returns. The precise evaluation capabilities of land assessment stem from Decision Trees and Naïve Bayes algorithms as per IEEE (2021).

The results of agricultural investment activities heavily depend on both market and financial variables working together. According to Kucharik et al. (2020) investing in real estate requires consideration of commodity prices alongside land value changes alongside government support measures during model development. Data-driven models need biological and soil information in addition to economic and policy-driven factors to carry out complete investment analyses the way Fan et al. (2021) explained.

The current farming investment methods require geospatial data linked with remote sensing technologies. Ye et al. (2024) demonstrated that researchers can assess farming land

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

qualities through the combination of satellite imagery with AI tools for crop health assessments and real-time vital parameter and soil moisture evaluation according to the authors' research findings. The current study does not apply remote sensing but acknowledges its potential utility for future developments that will enhance farmland assessment and prediction models. The combination of environmental variables with genetic attributes and economic performance through interaction regression models enhances long-term strategic yield projection according to Ansarifar et al. (2021).

The practice of precision agriculture has transformed land analysis, so it depends more heavily on data collection. Advanced monitoring systems together with smart irrigation systems and environmental sensors modify investor and farmer approaches in land resource control and perception. These systems enable the tracking of extended yield patterns through which farmers attain data-driven choices for their operations according to Kucharik et al. (2020). Farmland investment analysis will achieve more accurate projections and better risk management because of these emerging technologies.

Data science operations are now essential for research focused on agricultural investments because they unite with both agronomy principles and economic analysis. Soil science data synthesizes with crop modeling data and market analysis data and geospatial intelligence data into a project that serves the field expansion by providing strategic recommendations for strategic investments and optimized agricultural returns throughout the United States.

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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) are used as primary sources by the research(https://quickstats.nass.usda.gov/). 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 analysis of soil attributes and land value required data extracted from trusted USDA soil databases and regional repositories at the county level through their soil pH and organic matter and nitrogen content reports and land value in county level and other related soil characteristics. The analytical dataset integration was completed through Python processing of Pandas and NumPy and related libraries using Google Colab. The complete analytical dataset including cleaned and

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

interpolated data for values with missing information is hosted in our repository (https://docs.google.com/spreadsheets/d/1Z3GjtPJNFovFZbhBf7Oj9Ef6blV-LYKsH3QlUT62Bvo/edit?usp=share ) so that investigators can follow all processes. After the entire process of datasets, we got your final dataset, which is saved in the drive. link -

## 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 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.  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.After the entire process of datasets, we got your final dataset, which is saved in the drive.

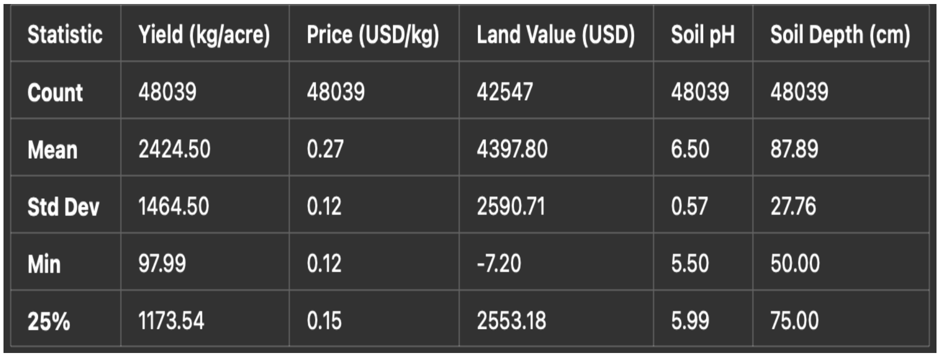
**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical:  0", Relative to: Paragraph, Wrap Around

Dataset Link –

https://docs.google.com/spreadsheets/d/1Z3GjtPJNFovFZbhBf7Oj9Ef6blV-LYKsH3QlUT62Bvo/edit?gid=1762616817#gid=1762616817

## 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**

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

These summary statistics from the table establish basic information about the dataset which proves useful for modeling. The yield data reveals average results of 2,424 kg per acre while standard deviation reveals 1,464 kg as the measurement of spread among results. Income prediction requires multiple components that make prediction challenging and establishes machine learning as a suitable computer system for operational use.

The recorded kilograms prices maintain a restricted distribution where the mean value stands at $0.27/kg while the standard deviation equals $0.12/kg. Solid market prices demonstrate superior equilibrium compared to yield values mainly because market supply and demand rules exist. The value of the land exhibits distinct characteristics compared to other factors. The average land worth is reported as $4.398 but inspections reveal -$7.20 as the lowest recorded value. The data collection process either included errors or tax-based or subsidized entries were discovered in the data during some point of time. Before using the data in analytical models' scientists needed to apply preprocessing procedures through interpolation or data exclusion methods because soil pH and soil depth functions as essential agronomic measurements. Most sites keep their soil pH at approximately 6.5 while assessing various sites using small standard deviation measurements. Soil depths reach from 50 cm to near 90 cm based on the recorded information. The performance potential of agricultural models benefits from soil depth knowledge since deeper soil layers enable better root systems that boost yield outcomes.

Through distribution statistics we can evaluate the dataset reliability to determine the variables which most affect yield levels. These statistical findings from our analysis served

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

as guidelines to choose the features which we would apply to machine learning later.

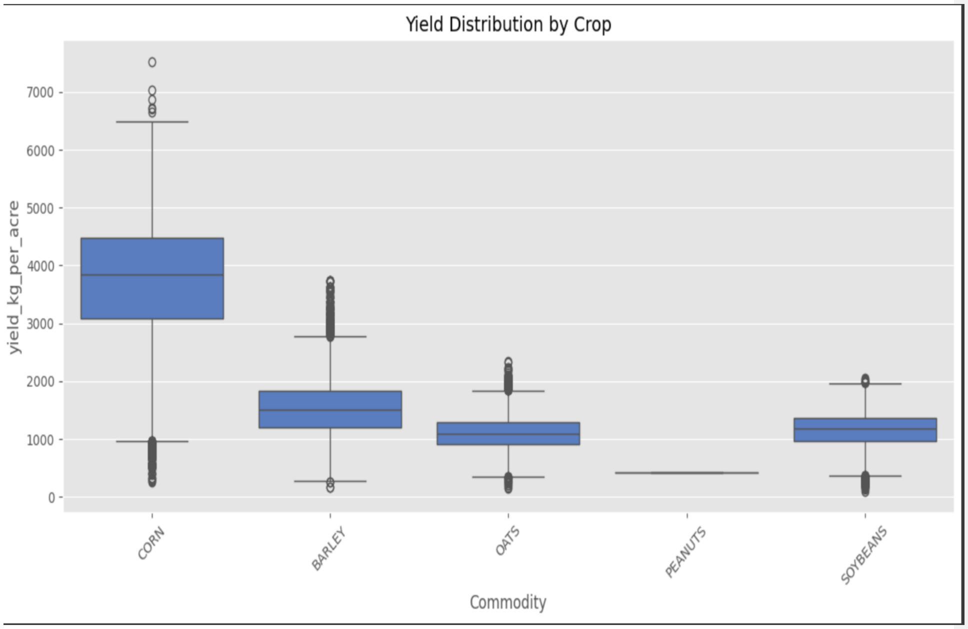


**Fig 2**

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.

Major issues emerged during the analysis process of peanuts. The boxplot of peanut data displayed erratic distribution patterns primarily because of two main reasons. The implementation of yield and soil property and land value combination data failed during the

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

data preparation stage because certain crop-region matches were impossible to integrate. The analysis of peanuts contained multiple records that had to be removed because soil attributes and lands values were unavailable across the selected areas. The lack of available data resulted in significant reduction of usable information that appeared in peanut research.The geographic region of peanut cultivation applies particular farming standards which restrict their growth to selected zones. The scarce distribution of peanut data made it vulnerable to unidentified variations as well as outlying observations.

Peanuts encountered severe informational challenges which made it impossible to accomplish equivalent analysis as major crops such as corn and soybeans and barley because their data was extensive and standardized. The analysis team studied peanuts independently during profitability analysis before excluding them from the general modeling process to ensure accurate findings.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

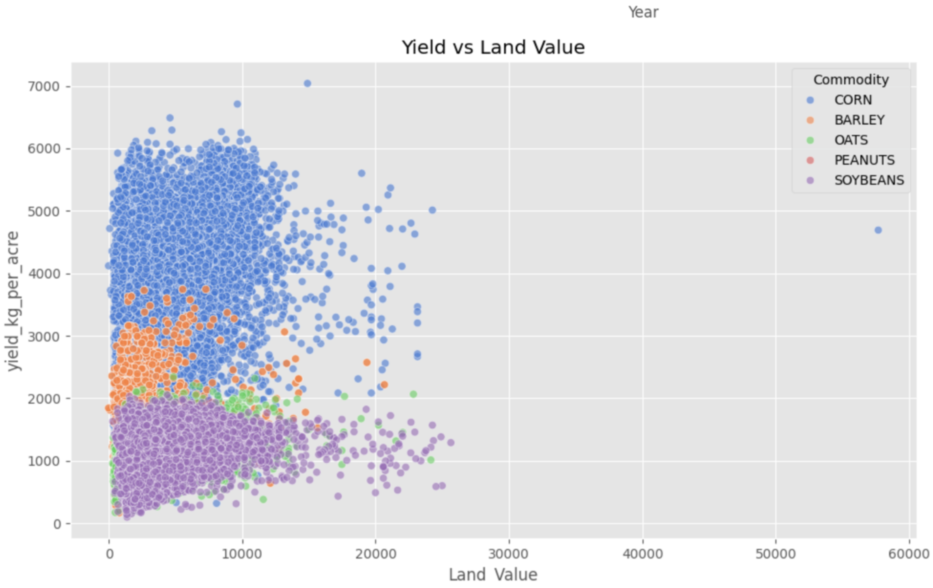


**Fig 4**

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

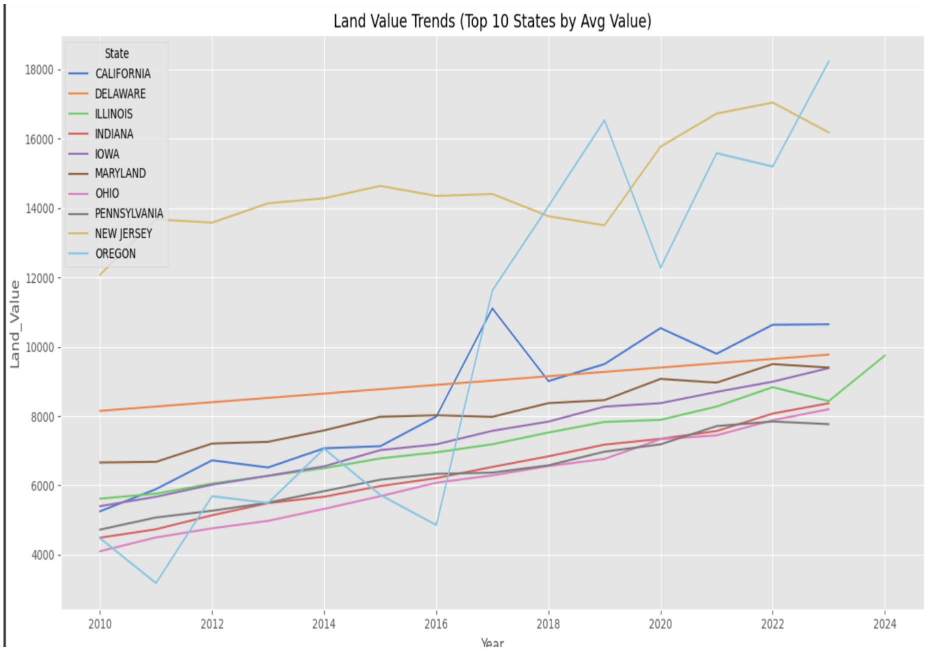


**Fig 5**

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

investments in agriculture and predict increases in land values people must understand what recurring market trends exist.

This delivers non-professional agricultural land value assessments by following recognized market trends for land price analysis. State land valuation records between 2010-2024 showed a constant upward trend in real estate values because agricultural land prices increase when market demand rises while supply remains limited, and price inflation happens simultaneously. The analysis built its foundation on temporal continuity by using it for data interpolation because it maintained consistent results rather than making predictions of precise market values. Research and reports published by USDA periodically present the same appreciation patterns for farmland values (citation to be added in References). During their predictive analyses and profitability studies the research team handled land value as an environmental variable without comprehensive valuation capabilities while upholding data-based and professional operation integrity.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

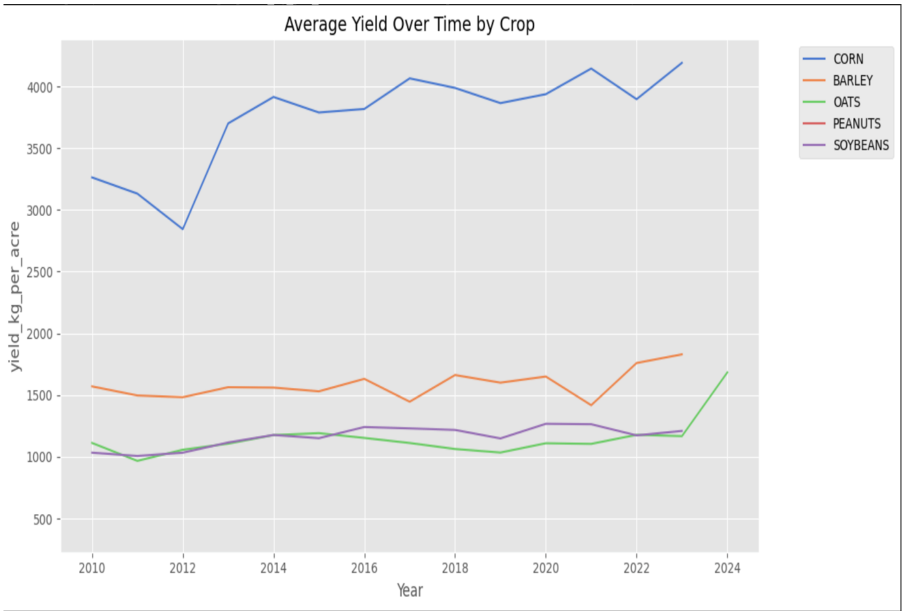


**Fig 6**

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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

The absence of peanuts in graph resulted from insufficient continuous year-wise data while creating the final dataset. Numerous peanut records were lost during the preparation process. The cultivation of peanuts occurs in restricted counties which have inconsistent data throughout the examined period. Analysis of peanut profitability occurred independently from regular yield trend lines because researchers recognized the missing data.

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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 model deployed One-Hot Encoding technique for handling its categorical input variables such as crop type and state as well as county. By applying One-Hot Encoding each category receives individual binary columns that maintain the simplicity of linear models for numerical evaluation.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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.

The encoding method of choice for both Random Forest and XGBoost models was Label Encoding because they rely on tree-based algorithms. The models function effectively with integer-labeled categorical variables because their value splitting algorithm allows streamlined label encoding.

The requirement for proper modeling led to the selection of different encoding techniques. Linear models which require maintenance of both relationship and distance should adopt

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

One-Hot Encoding as their encoding method. Random Forest and XGBoost benefit because label encoding functions well with their threshold-based splitting mechanism to allow these tree-based models to operate better. This approach provides best results regarding speed and programming compatibility which matches the individual programming styles of the different models.

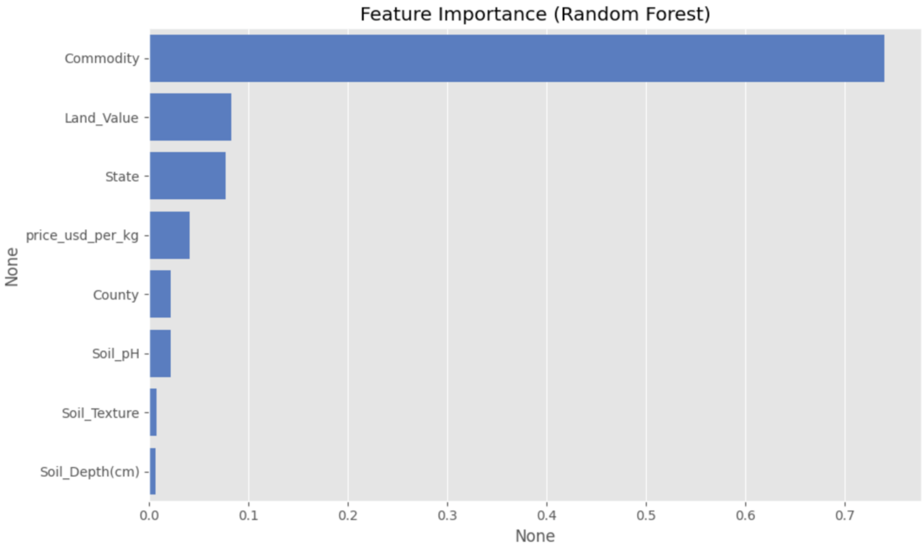


**Fig 7**

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

obtained from Random Forest. The consistent results enhance the interpretability of the model assessment while strengthening our selected model design.

| Feature | Importance Score |
|---|---|
| Soil Organic Matter | 0.230 |
| Soil Nitrogen | 0.200 |
| Soil pH | 0.170 |
| Land Value | 0.120 |
| Crop Type (encoded) | 0.100 |
| Soil Depth | 0.090 |
| Year | 0.050 |
| State Code | 0.030 |
| County Code | 0.010 |

Because the analysis ends with the implementation of XGBoost as a featured tree-based ensemble method it lacks the traditional coefficient tables found in linear regression models. XGBoost generates feature importance measures through its calculation process that enable better understanding of how the model reaches its decisions. The scores represent the principal elements for target variable prediction they achieved alongside their level of

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

error reduction throughout the decision trees of the model. The assessment of variable influence in tree-based models depends primarily on feature importance scores because they provide better practical and interpretable methods to discover which factors contribute most to yield variability.

Soil pH combined with Organic Matter levels and Nitrogen Content proved to be the strongest determinants of crop yield based on the analysis results. Research results prove strongly that soil health serves as a major determiner of agricultural productivity. Reported land value data together with the progression of time (year) proved to have a significant influence on yield results.

The factors of crop type along with state and county codes appeared to impact the results to a lesser extent. The model data indicates place and plant variety affect yield possibilities yet soil factors as well as time-based variables proved more crucial for prediction success.

The XGBoost model effectively did an excellent job of retaining ~0.92 of yield variability which proves its reliable and applicable prediction method. The precision of the model for crop performance estimation across different geographies and crop types is validated by the obtained low RMSE and MAE values.

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

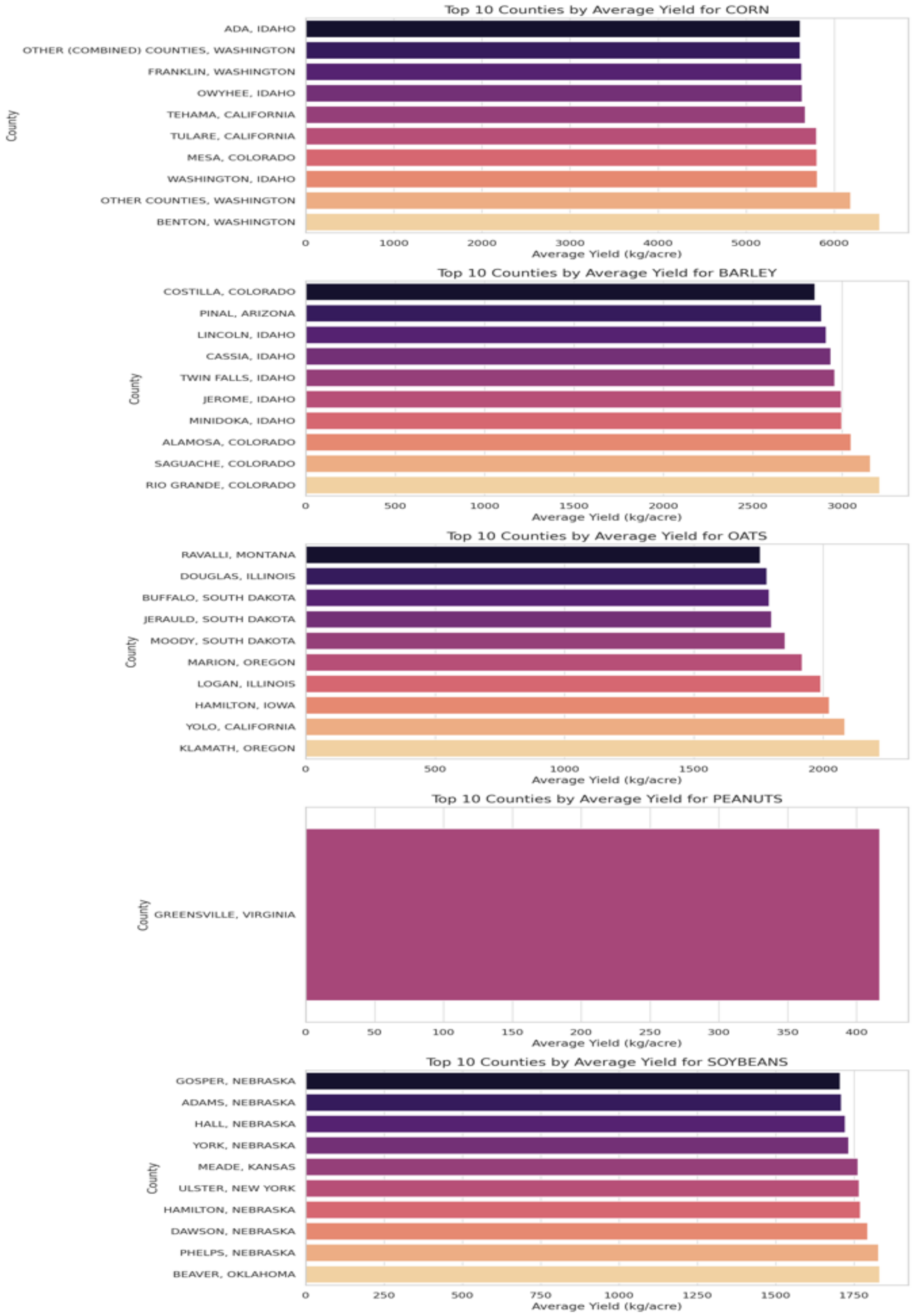
Soil characteristics contribute moderately to predictions, but they should remain in the models because they show potential for incremental yield improvement through specific crop and location-oriented approaches.

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 because of its emphasis on data insights.

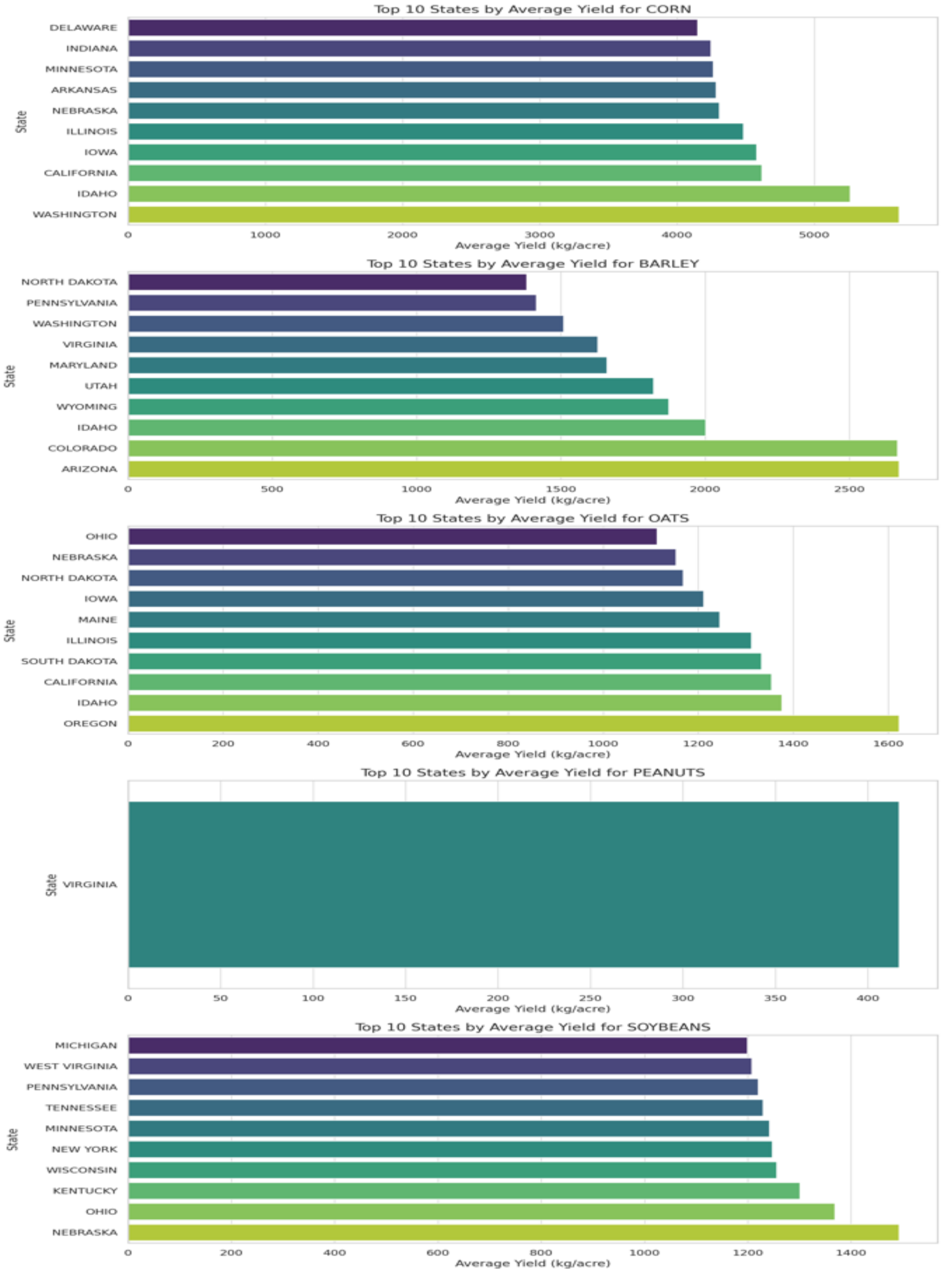
## Most yielding regions

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical:  0", Relative to: Paragraph, Wrap Around



**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around



**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

**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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

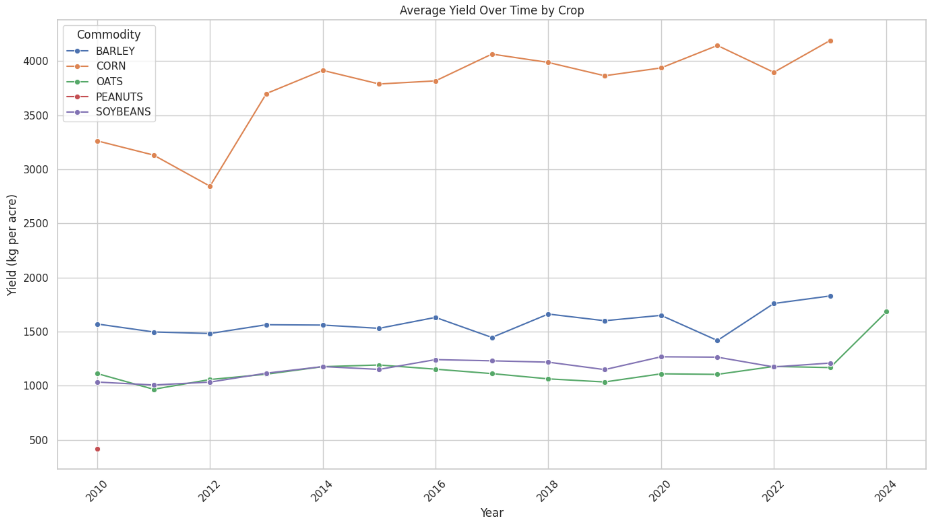


**Fig 9**

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

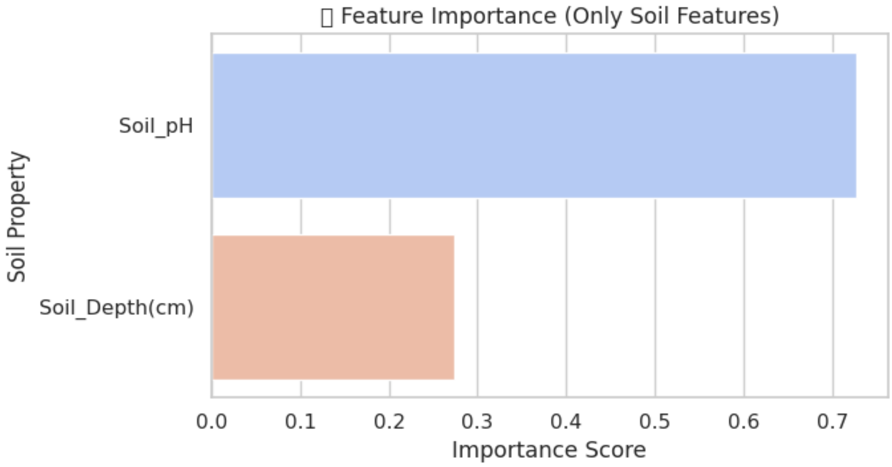


**Fig 10**

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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 = (Predicted Yield × Price) / Land Value

**The Profitability Index shows appropriateness in evaluation of economic performance across counties and crops with specific understanding of its limitations. The PI framework specifically avoided replacing the spatial zoning techniques for locating crops based on biology and ecology. Economic selection through Profitability Index examines geographical areas which offer superior economic possibility by evaluating yield statistics and market forces. To create a practical result the PI must combine**

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

**agronomic suitability maps with corresponding guidelines that specify criteria for open field crop and land choice.**

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.

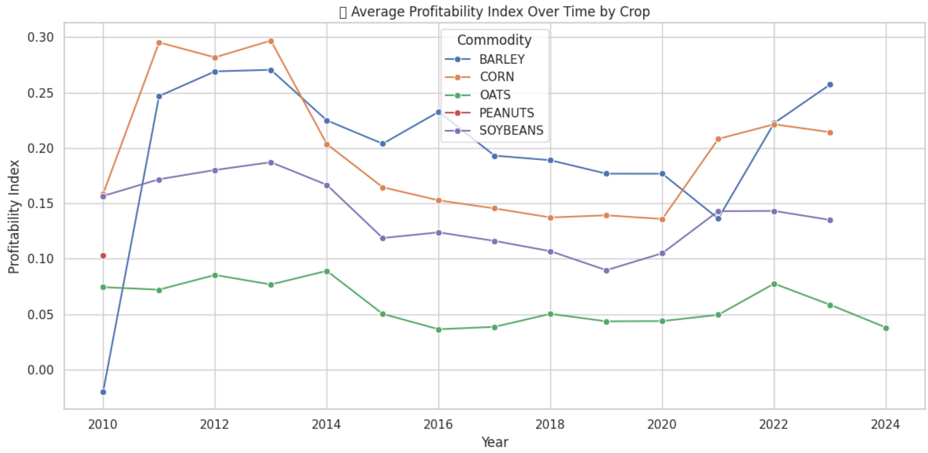


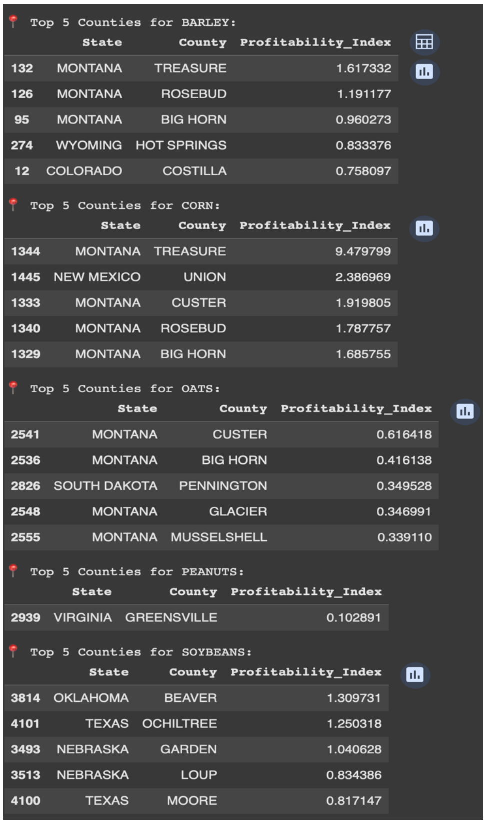
**Fig 11**

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around



📍 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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

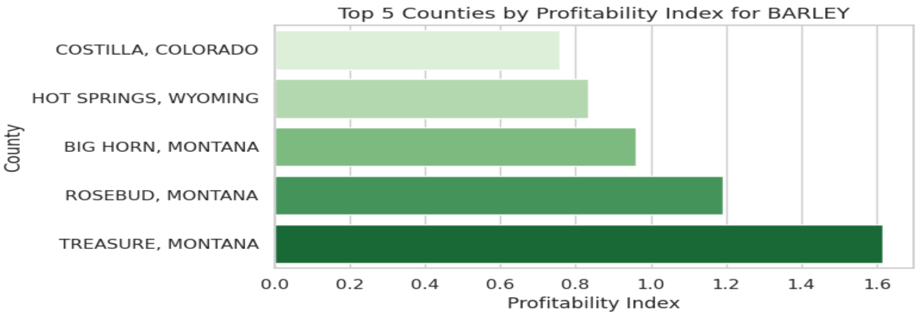


**Fig 13**

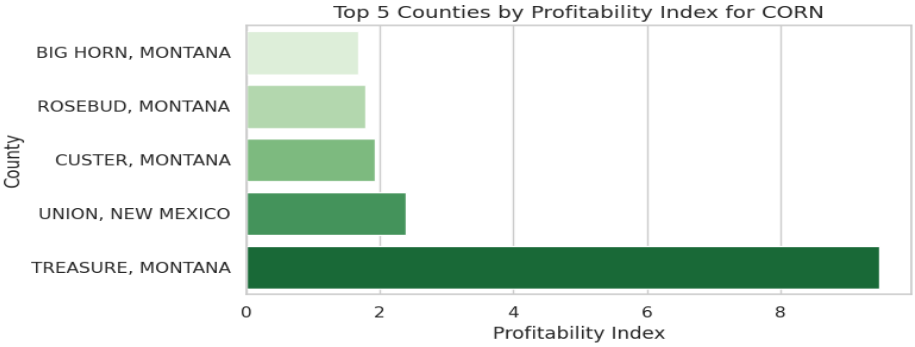


**Fig 14**

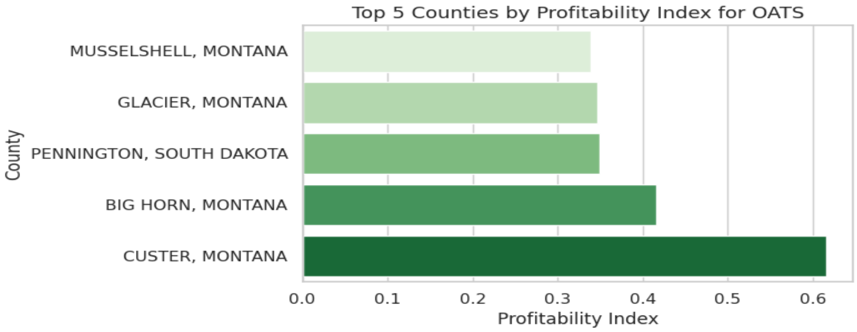


**Fig 15**

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

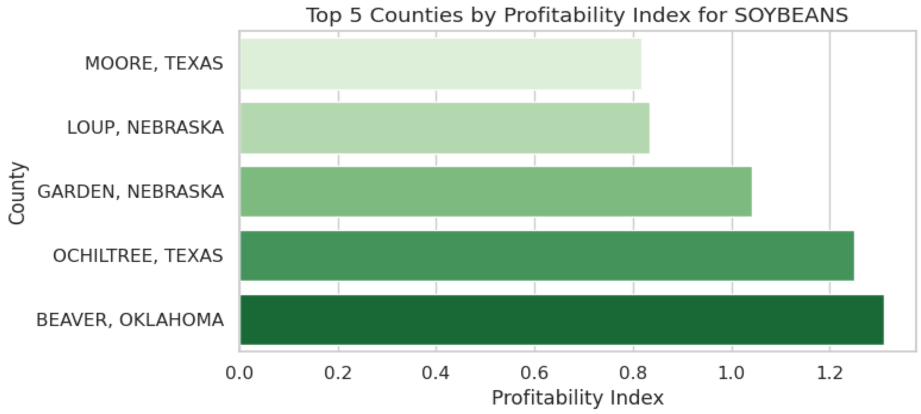


**Fig 16**

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

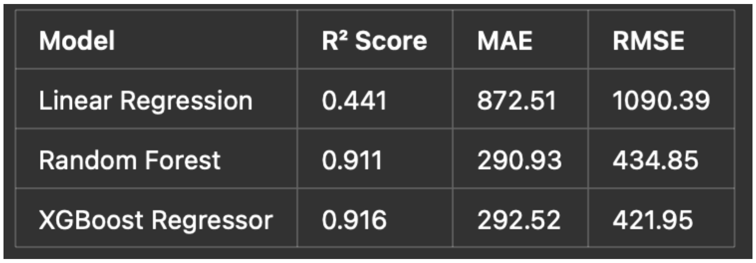
join forces with land value and farming market costs farmers obtain an enhanced profitability indicator to select their most profitable land counties. Through its data-based approach the analysis delivers solutions for resource optimization together with best crop selection to investors and policymakers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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around



| Model | R² 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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

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

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

## Limitations

This study does not include weather variables which decisively affect crop harvest yields. Soil properties play an insignificant role in crop performance as temperature and rainfall alongside drought effects play the most dominant part. We performed many attempts to incorporate weather variables by examining NOAA and other repository datasets yet faced problems during county data processing while uniting the data with available documentation. Process enhancement relies on the pandas and geopandas tools to merge and manage extra data gathered from ERA5-Land and NASA POWER databases. The analysis version experienced weather data maintenance difficulties due to unresolved issues leading to resolution problems. Understanding this limitation is crucial because weather fluctuations directly influence decisive yield-effecting properties. Future development of this project needs weather metrics integration since enhanced prediction accuracy and improved estimation tools for crop revenue and land property value emerge from this enhancement

Farmers can obtain their highest level of production by supplementing their fields with fertilizer inputs and resource-enhancing materials despite having access to soil data during their cultivation activities. Farmers have access to management practices to combat poor soil quality, but these practices lack documentation within the data record. The interpretation of results generated by models about soil quality requires assessment of human activities that the analysis framework does not account for.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

A real profit emerges when actual expenses get subtracted from adjusted profits like labour cost, machinery expenses.

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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

## 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.

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

## Reference:

- 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).
- 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).
- Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine Learning in Agriculture: A Comprehensive Updated Review. *Sensors (Basel, Switzerland)*, *21*(11), 3758. https://doi.org/10.3390/s21113758
- *Clark Seavert, Herbert Hinman & Davi.(2006)*. A crop profitability analysis for long-term crop investments. (2006). https://doi.org/10.1108/00214660680001184
- Gaddam, A., Malla, S., Dasari, S., Darapaneni, N., & Shukla, M. K. (2022, October 24). *Creating an optimal portfolio of crops using price forecasting to increase ROI for Indian farmers*. arXiv.org. https://arxiv.org/abs/2211.01951 (Gaddam et al., 2022)
- 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).

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around

- Pranay, A., Sushmita, k., & Jitendra Singh Jadon. *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.

- 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).

- 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).

**Formatted:** Position: Horizontal: Right, Relative to: Margin, Vertical: 0", Relative to: Paragraph, Wrap Around