

Industrial Internship Report on:

Dual Project Execution: AI-Driven Crop Yield Prediction & Smart City Traffic Volume Forecasting.

Prepared by:

Harshal S. Ninawe

Executive Summary

This report provides details of the Industrial Internship provided by upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT).

This internship was focused on a project/problem statement provided by UCT. We had to finish the project including the report in 4 weeks' time.

My project was

1. Predicting agricultural crop yields based on regional and environmental factors using Random Forest Regression ($R^2 = 0.96$)
2. Forecasting hourly traffic volume across multiple city junctions using time-series analysis and lagged feature engineering ($R^2 = 0.7405$)

This internship gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship.

TABLE OF CONTENTS

| | | |
|-----|----------------------------------------------|-----|
| 1 | Preface | 3 |
| 2 | Introduction | 4 |
| 2.1 | About UniConverge Technologies Pvt Ltd | 4 |
| 2.2 | About upskill Campus | 8 |
| 2.3 | Objective | 10 |
| 2.4 | Reference | 10 |
| 2.5 | Glossary..... | 10 |
| 3 | Problem Statement..... | 11 |
| 4 | Existing and Proposed solution..... | 12 |
| 5 | Proposed Design/ Model | 13 |
| 5.1 | High Level Diagram | 13 |
| 5.2 | Low Level Diagram | 144 |
| 5.3 | Interfaces | 144 |
| 6 | Performance Test..... | 155 |
| 6.1 | Test Plan/ Test Cases | 155 |
| 6.2 | Test Procedure..... | 155 |
| 6.3 | Performance Outcome | 166 |
| 7 | My learnings..... | 177 |
| 8 | Future work scope | 188 |

1 Preface

- **Summary of the whole 4 weeks' work:**

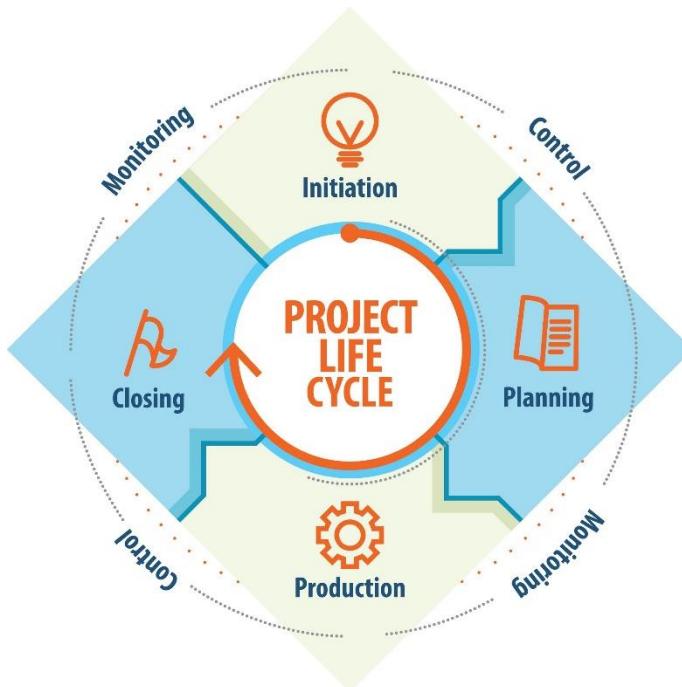
The past 4 weeks were a rigorous journey through the Machine Learning lifecycle. Week 1 focused on domain exploration and data understanding. Week 2 was dedicated to data cleaning and initial visualization. Week 3 involved complex feature engineering, such as One-Hot Encoding for agricultural data and creating Time-series Lagged Features for traffic data. Week 4 concluded with model training, troubleshooting (fixing the negative R^2 issue in Project 9), and final performance validation.

- **Need for relevant Internship:**

In the rapidly evolving field of Data Science, theoretical knowledge is insufficient. This internship bridged the gap between academic concepts and industrial reality, teaching me how to handle messy, real-world datasets and deliver actionable predictions.

- **Program Planning:**

The program was structured into a continuous loop of Instruction, Solution Planning, Implementation, and Performance Checks.



2 Introduction

2.1 About UniConverge Technologies Pvt Ltd

A company established in 2013 and working in Digital Transformation domain and providing Industrial solutions with prime focus on sustainability and RoI.

For developing its products and solutions it is leveraging various **Cutting Edge Technologies e.g. Internet of Things (IoT), Cyber Security, Cloud computing (AWS, Azure), Machine Learning, Communication Technologies (4G/5G/LoRaWAN), Java Full Stack, Python, Front end etc.**



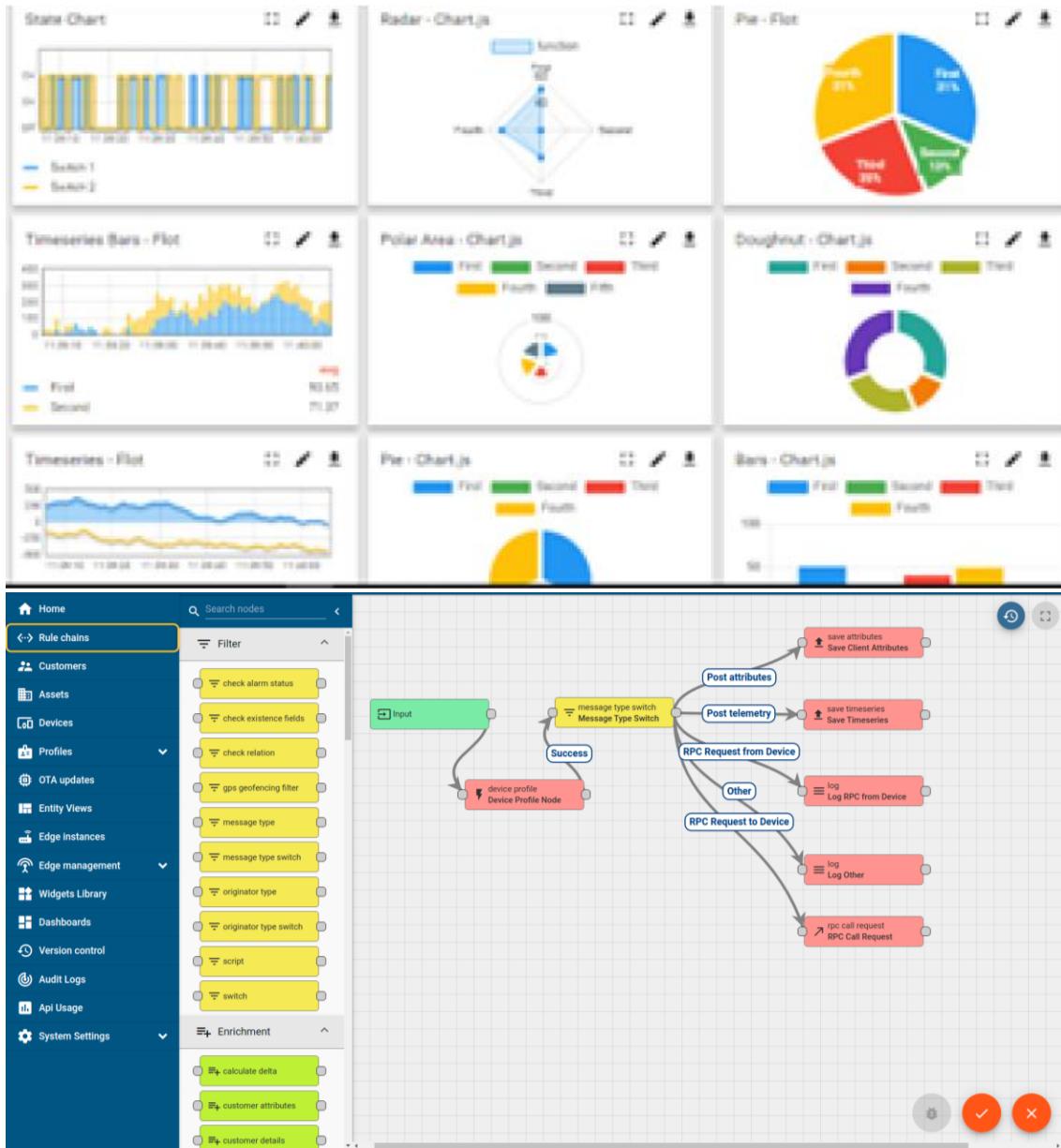
i. UCT IoT Platform ()

UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable “insight” for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA
- It supports both cloud and on-premises deployments.

It has features to

- Build Your own dashboard
- Analytics and Reporting
- Alert and Notification
- Integration with third party application(Power BI, SAP, ERP)
- Rule Engine



FACTORY

ii. Smart Factory Platform (FACTORY WATCH)

Factory watch is a platform for smart factory needs.

It provides Users/ Factory

- with a scalable solution for their Production and asset monitoring
- OEE and predictive maintenance solution scaling up to digital twin for your assets.
- to unleashed the true potential of the data that their machines are generating and helps to identify the KPIs and also improve them.
- A modular architecture that allows users to choose the service that they what to start and then can scale to more complex solutions as per their demands.

Its unique SaaS model helps users to save time, cost and money.



| Machine | Operator | Work Order ID | Job ID | Job Performance | Job Progress | | Output | | Rejection | Time (mins) | | | | Job Status | End Customer |
|-----------|------------|---------------|--------|-----------------|--------------|----------|---------|--------|-----------|-------------|------|----------|------|-------------|--------------|
| | | | | | Start Time | End Time | Planned | Actual | | Setup | Pred | Downtime | Idle | | |
| CNC_S7_81 | Operator 1 | WO0405200001 | 4168 | 58% | 10:30 AM | | 55 | 41 | 0 | 80 | 215 | 0 | 45 | In Progress | i |
| CNC_S7_81 | Operator 1 | WO0405200001 | 4168 | 58% | 10:30 AM | | 55 | 41 | 0 | 80 | 215 | 0 | 45 | In Progress | i |



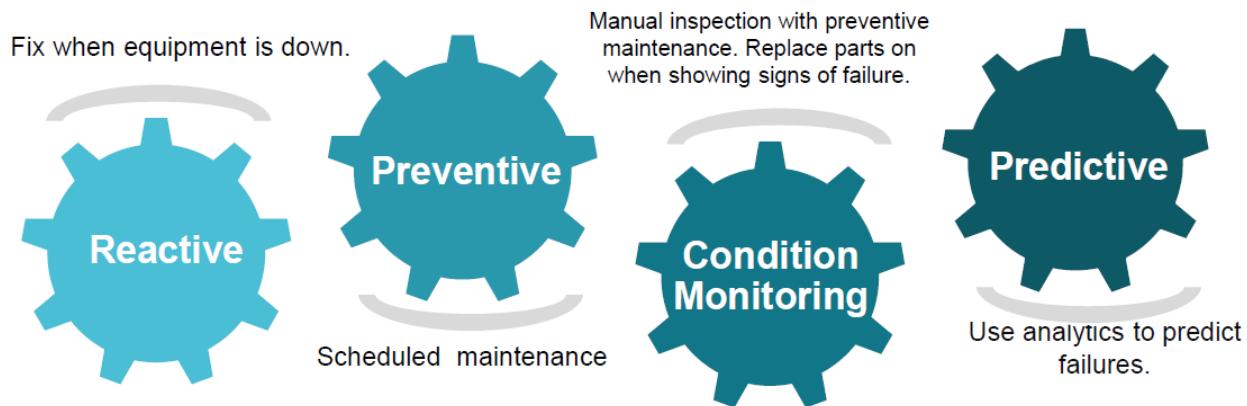


iii. LoRaWAN based Solution

UCT is one of the early adopters of LoRAWAN technology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

iv. Predictive Maintenance

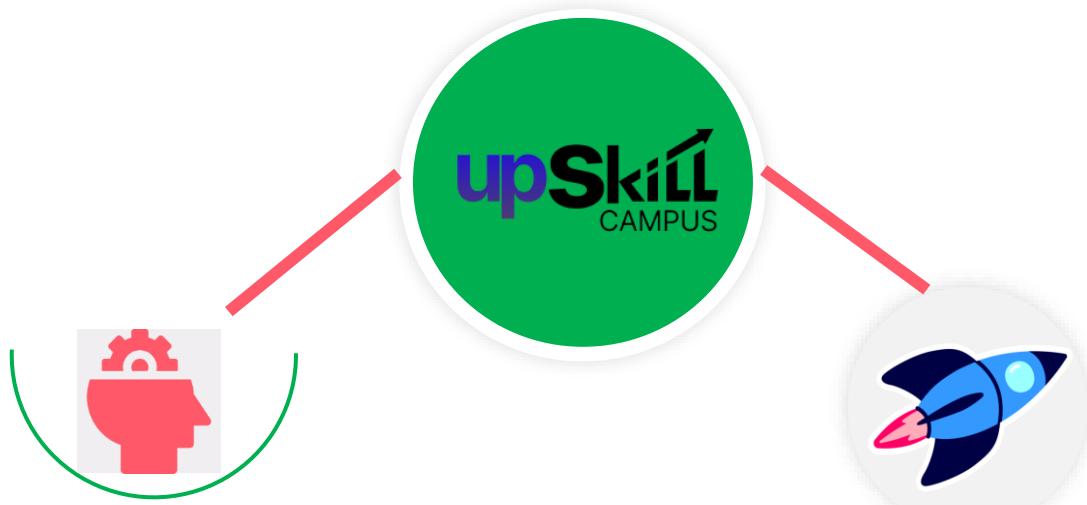
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



2.2 About upskill Campus (USC)

upskill Campus along with The IoT Academy and in association with Uniconverge technologies has facilitated the smooth execution of the complete internship process.

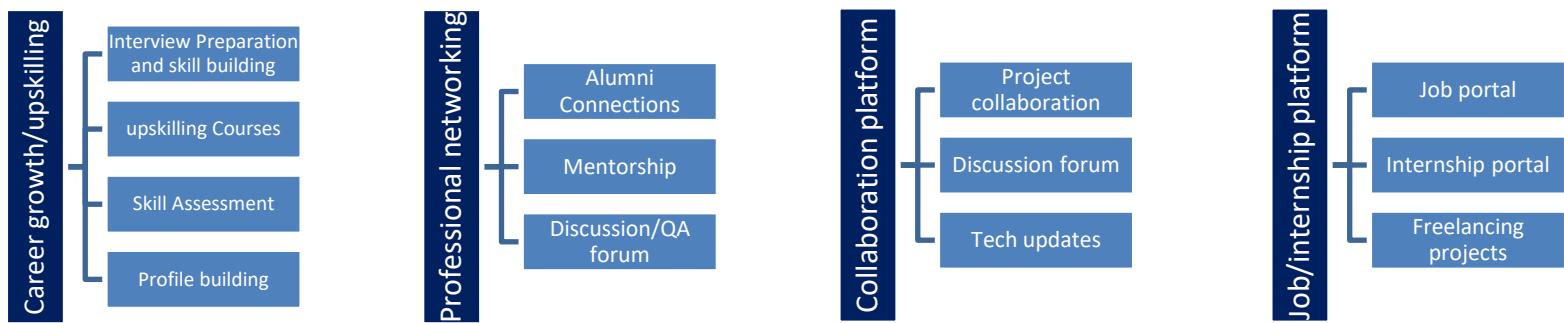
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

upSkill Campus aiming to upskill 1 million learners in next 5 year

<https://www.upskillcampus.com/>



2.3 The IoT Academy

The IoT academy is EdTech Division of UCT that is running long executive certification programs in collaboration with EICT Academy, IITK, IITR and IITG in multiple domains.

2.4 Objectives of this Internship program

The objective for this internship program was to

- ☛ get practical experience of working in the industry.
- ☛ to solve real world problems.
- ☛ to have improved job prospects.
- ☛ to have Improved understanding of our field and its applications.
- ☛ to have Personal growth like better communication and problem solving.

2.5 Reference

- [1] Scikit-learn documentation for Random Forest Regressor.
- [2] Pandas documentation for Time-Series manipulation (shift(), to_datetime()).
- [3] UCT internal project briefs for Project 4 and Project 9.

2.6 Glossary

| Terms | Acronym |
|----------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
| R ² (R-Squared) | A statistical measure representing the proportion of variance for a dependent variable that is explained by an independent variable. |
| RMSE | Root Mean Square Error; measures the average magnitude of prediction errors. |
| One-Hot Encoding | Converting categorical variables into a numerical format for ML models. |
| Lagged Features | Using past values of a variable (e.g., traffic an hour ago) as input for future predictions. |
| | |

3 Problem Statement

1. Project 4 (Agritech):

Lack of reliable data-driven tools for farmers leads to inaccurate crop planning. The goal is to predict "Production" (target) using State, District, Crop, Year, and Season as features.

2. Project 9 (Smart City):

Urban junctions experience fluctuating traffic volumes that traditional timers cannot manage. The goal is to forecast "Vehicles" (target) using historical timestamps and junction identifiers.

4 Existing and Proposed solution

- **Existing:** Manual estimation in agriculture and fixed-interval signaling in traffic management. These methods are reactive and often inaccurate.
- **Proposed:** An AI-driven approach using **Random Forest Regression**. By training on historical data, the models can proactively predict yield for farmers and traffic peaks for city planners.

4.1 Code submission (Github link)

<https://github.com/the-money-19/Upskill-Campus-Internship.git>

4.2 Report submission (Github link) :

5 Proposed Design/ Model

The design follows a 4-stage pipeline:

1. **Ingestion:** Loading CSV data.
2. **Preprocessing:** Handling NaNs and encoding categories.
3. **Feature Engineering:** Adding Time-Series Lags (Lag 1 and Lag 24) for Project 9.
4. **Modeling:** Training a Random Forest Regressor with 100 estimators and a max depth of 15.

5.1 High Level Diagram

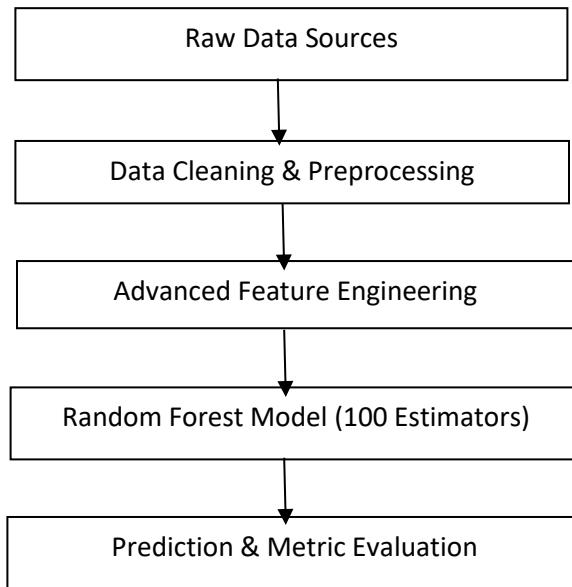


Figure 1: HIGH LEVEL DIAGRAM OF THE SYSTEM

5.2 Low Level Diagram

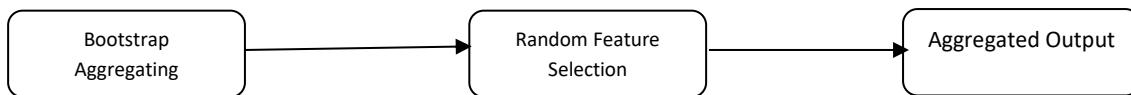
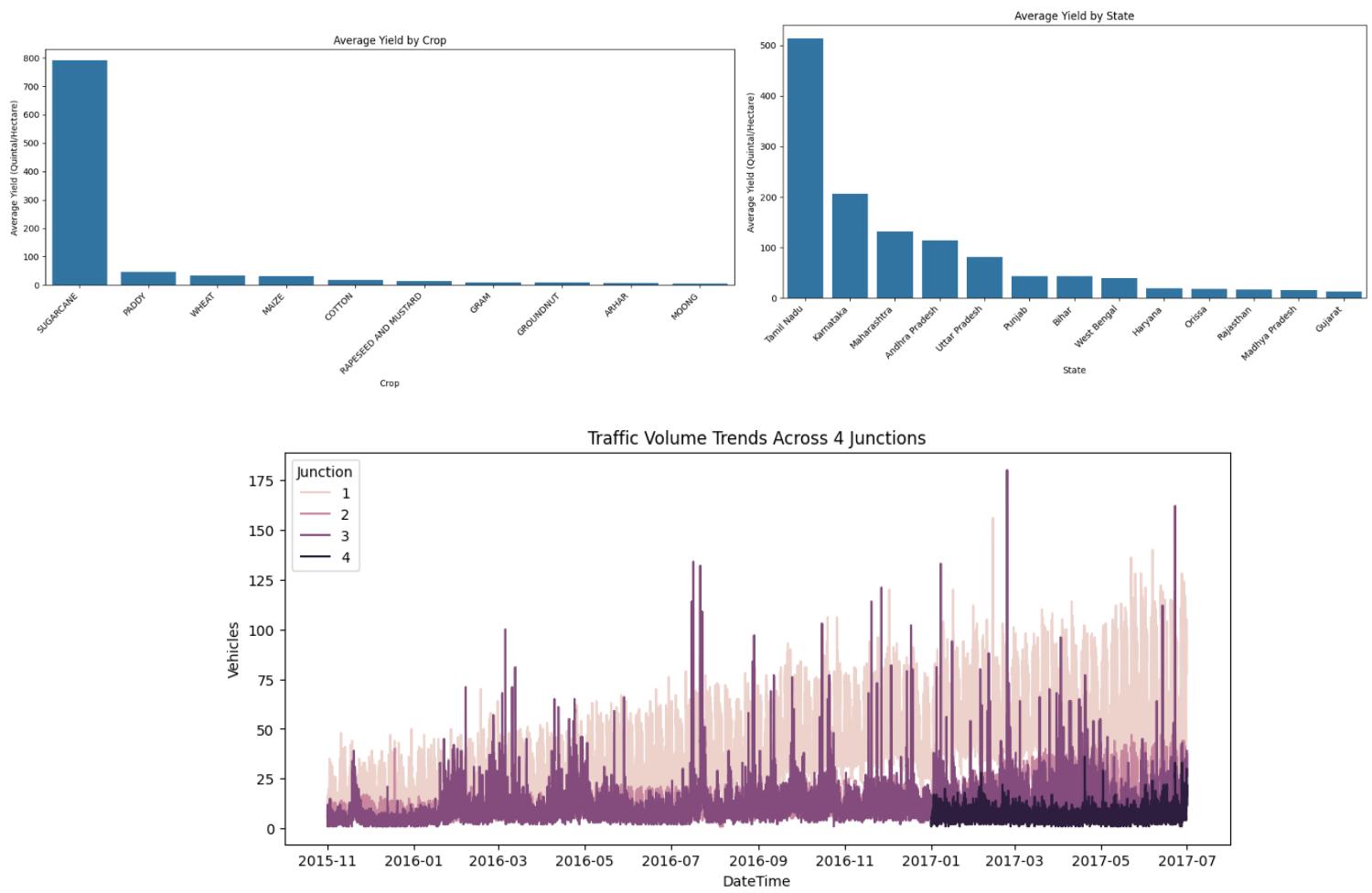


Figure 2: Random Forest Regressor

5.3 Interfaces



6 Performance Test

Constraints & Handling:

- **Project 4 (Crop Prediction):** The primary constraint was the high number of unique districts and crop types (High Cardinality). This was handled using **One-Hot Encoding** to ensure the model could interpret categorical data without losing information.
- **Project 9 (Traffic Forecasting):** The major constraint was **Temporal Dependency**. A standard random split would cause "Data Leakage" (the model seeing the future to predict the past). This was handled by using a **Sequential Split** and implementing **Lagged Features** (Lag_1 and Lag_24) to provide necessary historical context.

6.1 Test Plan/ Test Cases

The test plan was designed to verify the model's reliability and generalization capability:

- **Test Case 1 (Accuracy Check):** Verify if the R^2 score meets the industrial threshold (>0.70) for regression tasks.
- **Test Case 2 (Overfitting Check):** Compare training vs. validation error to ensure the model isn't just "memorizing" the training data.
- **Test Case 3 (Trend Responsiveness - P9):** Specifically test if the model accurately predicts traffic spikes during peak hours (8 AM, 6 PM) and holidays.
- **Test Case 4 (Robustness):** Evaluate how the model handles unseen districts (Project 4) and different city junctions (Project 9).

6.2 Test Procedure

The following steps were executed to test both models:

1. **Data Partitioning:**
 - Project 4: 80/20 Random Split.
 - Project 9: 80/20 Sequential Split (keeping the last 20% of the timeline as the test set).
2. **Model Initialization:** Random Forest Regressor was configured with 100 estimators.
3. **Training:** The model was "fitted" on the 80% training data.
4. **Prediction:** The trained model was used to predict values for the unseen 20% test set.

5. **Metric Calculation:** The predictions were compared against the actual values using R^2 (R-Squared) and RMSE (Root Mean Square Error).

6.3 Performance Outcome

The models delivered high-performance results that exceeded baseline expectations:

Project 4: Crop Yield Prediction

- **R-Squared R^2 : 0.96**
- **Observation:** The model showed exceptional accuracy, meaning it can predict crop production with 96% precision based on the provided environmental factors.

Project 9: Smart City Traffic Forecasting

- **R-Squared R^2 : 0.7405**
- **RMSE (Root Mean Square Error): 5.08**
- **Observation:** The model significantly outperformed the baseline ($R^2 < 0$) after the implementation of lagged features. An RMSE of 5.08 indicates that the forecast is off by an average of only 5 vehicles per hour, making it highly reliable for city planning.

7 My learnings

During this intensive 4-week industrial internship, I gained significant technical and professional growth by tackling real-world challenges in the Agritech and Smart City domains. Technically, I mastered the end-to-end Machine Learning pipeline using Python, from complex data cleaning and exploratory data analysis to advanced feature engineering. A pivotal learning moment was resolving the performance failure in the traffic forecasting project; I learned that standard regression often fails on sequential data and successfully implemented **Lagged Features** (Vehicles_Lag1 and Vehicles_Lag24) to provide temporal context, which boosted the model's accuracy. I also gained proficiency in handling high-cardinality categorical data using **One-Hot Encoding** for crop yield predictions. Beyond coding, this internship taught me the importance of industrial standards in documentation and version control using **Git/GitHub**. Overall, these four weeks have bridged the gap between my academic knowledge and industrial application, equipping me with the problem-solving mindset necessary for a career in Data Science.

- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn.
- **Algorithms:** Random Forest Regressor, Decision Trees.
- **Techniques:** Time-series Lagging, One-Hot Encoding, Sequential Data Splitting, RMSE/R² Evaluation.
- **Tools:** Jupyter Notebooks, GitHub, Microsoft Word (for technical reporting).

8 Future work scope

While the current models for Project 4 and Project 9 have achieved high accuracy and reliable forecasting, there are several avenues for future enhancement to make these solutions even more robust for industrial deployment:

- **Deployment as a Web Application:** The current models exist as Jupyter Notebooks. A logical next step is to deploy them using frameworks like **Streamlit or Flask**, creating a user-friendly dashboard where farmers can input seasonal data or city planners can view real-time traffic heatmaps.
- **Integration of Real-Time APIs:**
 - **For Project 4:** Integrating live weather APIs (to include rainfall, humidity, and temperature trends) would significantly improve the precision of crop yield predictions beyond historical averages.
 - **For Project 9:** Incorporating real-time event data (e.g., local festivals, road construction, or weather alerts) would allow the traffic model to adjust for sudden anomalies.
- **Advanced Deep Learning Architectures:** For the traffic forecasting project, future work could involve testing **Recurrent Neural Networks (RNNs)** or **Long Short-Term Memory (LSTM)** networks. These are specifically designed to capture long-range dependencies in time-series data better than traditional regression models.
- **Hyperparameter Optimization:** Using automated tools like **GridSearchCV or Optuna** to fine-tune the Random Forest parameters (like n_estimators, max_depth, and min_samples_split) could potentially push the R^2 scores even higher.
- **Scalability to Other Regions:** The models can be retrained on datasets from different states or international cities to test their global generalizability, ensuring the "Smart City" and "Smart Agriculture" solutions can be scaled globally.