QueueML

An ML Based Application for Waiting Time Prediction and Queue Management

Anjale Ramesh

Date: 05/06/2025

1. Problem Statement

In service-based environments such as hospitals, banks, airports, and call centres, customers, calls, or patients typically arrive at a service facility with the expectation of being served. If service is unavailable at the moment of arrival, they must wait in line until it's their turn to be served. Efficiently managing such waiting lines is crucial for both operational efficiency and customer satisfaction.

When queues become excessively long, it causes significant problems for the service system. Not only does this reduce the overall efficiency of the system, but it also leads to customer frustration and dissatisfaction. Long queues are a clear sign of congestion, which can have serious consequences, including delays, overcrowding, and a negative customer experience. The motivation of this study is to better control this congestion in order to optimize the service system. Congestion in the system can be primarily measured in two key ways: the number of customers in the queue and the waiting time before an entity receives service. These two factors are often interlinked. For example, a large number of customers in the queue typically leads to longer waiting times. If an arriving customer is able to know how long they will need to wait, it can improve their experience, as it helps manage their expectations and allows them to plan accordingly.

To address these challenges, there's a clear need for a smart, adaptive system that can both predict and optimize waiting times. This would allow service providers to more effectively manage customer flow, reduce waiting times, and enhance customer satisfaction. By leveraging both historical data (such as past service and arrival patterns) and real-time data (such as current queue status), a system could provide real-time predictions of waiting times and suggest adjustments to improve service efficiency.

QueueML is a machine-learning-based software tool developed to address this gap. It is designed to predict waiting times in service systems by analysing historical customer arrival and service data. The tool uses machine learning techniques to identify patterns in the data, forecast queue lengths, and predict the time customers will have to wait before receiving service. Additionally, it offers optimization strategies to adjust queue management in real-time, helping service providers minimize waiting times and avoid congestion.

QueueML is aimed at small and medium-sized businesses in various sectors, enabling them to improve their operational efficiency without the need for costly infrastructure. By leveraging the power of machine learning, this tool helps businesses offer better customer experiences through smarter queue management, thereby improving customer satisfaction and retention.

1.1 Initial Needs Statement

There is a need for an affordable, intelligent tool that can help small and medium service providers manage queues and reduce waiting times by predicting peak hours and suggesting optimal staff deployment strategies.

2. Market and Customer Needs Assessment

2.1 Market Overview:

Small/medium businesses form the backbone of the service sector in many developing and emerging economies. These businesses often include outpatient clinics, diagnostic labs, local government centres, educational institutions, and customer service desks. Despite handling a high volume of customers, these establishments often operate with minimal digital infrastructure and lack access to advanced queue management solutions. They require cost-effective, easy-to-use solutions to:

- A reliable prediction of wait times to manage customer expectations.
- Clear visibility into peak traffic hours to plan staffing accordingly.
- Insights into service duration patterns to reduce bottlenecks.
- Improve customer satisfaction and retention

The cost of existing queue management systems is a major barrier for these service systems. Most available solutions are designed for large corporations and require expensive hardware, subscription fees, or technical expertise that smaller organizations cannot afford. A product like QueueML, with its focus on simplicity, affordability, and machine learning-driven insights, directly addresses these gaps and is well-positioned to deliver value to this segment.

2.2 Customer Needs:

Customer Segments Identified:

1. Outpatient clinics and diagnostic labs: Frequently experience rush hours during early mornings and evenings. Struggle to manage patient inflow and staffing schedules.
2. Small banks and government offices: Serve a diverse population, often face peak-hour congestion, especially at the beginning and end of the month.
3. Customer care service centres: Need systems to handle inquiries, complaints, and one-on-one counselling.

Customers need a platform that shows their queue position and the average expected waiting time to receive a particular service from the point of time they join the queue. With accurate wait time predictions displayed at entry points, customers can make decisions about whether to wait, reschedule, or return later. This reduces uncertainty, frustration, and perceived unfairness. In some cases, they may even receive updates or alerts if wait times change. Ultimately, QueueML empowers customers with timely information and improves their satisfaction with the service experience.

3. External Search

Extensive external research is conducted to understand the existing tools which provides queue management effectively.

3.1 Benchmarking

Qminder and QLess are both commercial queue management systems, but they are mainly designed for large organizations or enterprises.

Qminder

- Type: Cloud-based queue management and customer service system
- Used by: Large hospitals, government offices, banks, and retail
- Key Features:
    - Allows customers to sign in remotely or at a kiosk
    - Staff can manage queues from a dashboard
    - Offers real-time data and performance tracking
- Limitations:
    - More focused on ticketing and user interface
    - Expensive for small setups
    - No built-in ML/AI-driven wait time prediction

QLess

- Type: Virtual queueing and appointment scheduling system
- Used by: Universities, DMVs, healthcare systems
- Key Features:
    - Customers join a virtual queue via mobile
    - Real-time updates via SMS
    - Data analytics for business owners
- Limitations:
    - Subscription cost is high
    - Geared toward institutions with large visitor volumes
    - Focuses more on queue notification and appointment booking, less on data- driven optimization

The existing tools like Qminder and QLess focus on enterprise markets and are cost-prohibitive. No affordable ML-based queue optimization tools exist for small and medium businesses like QueueML.

3.2 Applicable Constraints

QueueML is developed with the practical realities of small and medium service providers in mind. These businesses typically operate with limited budgets, infrastructure, and historical data, which presents several constraints that must be addressed during product development:

- Cost: The product must be affordable and cost-effective. Should not demand expensive software licenses or high subscription fees.
- Data Availability: Many service systems do not maintain comprehensive digital records of customer arrivals and service times. The system must be designed to function effectively with limited historical data.
- Infrastructure: Every service provider may not have advanced IT infrastructure. QueueML must be lightweight and able to run on basic hardware like standard desktops or tablets.

These constraints shape QueueML into a lean, adaptable solution that aligns with the operational capabilities of small and medium service providers.

4. Product Prototype

QueueML works by collecting historical customer flow data, training an ML model to forecast expected waiting times, and displaying insights through a dashboard. Admins can log in, upload data or integrate with existing systems, and receive real-time queue insights and recommendations.

Abstract Product Design

Inputs:

- Customer arrival times (timestamped logs)
- Service time duration (per staff or service type)
- Staff schedules
- Day of week, public holidays, etc.

System Components:

- Data ingestion and preprocessing module
- ML engine for prediction
- Dashboard for visualization and alerts

Outputs:

- Real-time queue status
- Wait time forecasts

The interface will be simple, enabling staff to upload data or integrate basic appointment tools. The dashboard, built using tools like Streamlit, will visualize wait times and generate insights.

5. Product Details

5.1 How does it work?

There are different options to convey the details to the beneficiaries:

Option A: On a Digital Display at the Service Centre

- A monitor or tablet is placed in the waiting area (like a queue token screen).
- QueueML pushes real-time wait time estimates and position-in-queue data to this screen.
- Customers entering the centre can immediately see their expected waiting time.

Option B: Via a Mobile Link or SMS

- On check-in, the registered customer receives a link or SMS (from the centre’s system integrated with QueueML).
- The link opens a simple mobile page showing:
    - "You are #4 in line"
    - "Estimated waiting time: 12 minutes"

Option C: Printed Token with Wait Time

- If integrated with a token printer (used in many clinics/labs), the token includes:
    - Token number
    - Estimated wait time (powered by QueueML)

5.2 Data Sources

- **Internal Databases/Systems:** Data generated and maintained within the organization's core operational systems.
- **User/Customer Data:** Information directly related to the individuals who interact with the organization's products, services, or systems.
- **Content/Media Data:** Specifically refers to visual recordings captured by Closed-Circuit Television (CCTV) systems.

5.3 Algorithms and Frameworks Used

Machine Learning Algorithms Used:

- Random Forest Regression/ gradient boosting machine (GBM)/ XGBoost: Predict average wait time based on historical patterns
- Long Short-Term Memory (LSTM) networks: For time series modelling of customer arrivals due to their ability to learn and retain long-term dependencies in sequential

data.

Libraries/Frameworks: Pandas, NumPy, scikit-learn, TensorFlow/Keras, PyTorch

6. ML Model Development

The first focus is on predicting average waiting times using the supervised learning algorithm, random forest regression, and the second addresses the temporal patterns of customer arrivals using Long Short-Term Memory (LSTM) networks.

Predicting Waiting Time with Supervised Learning

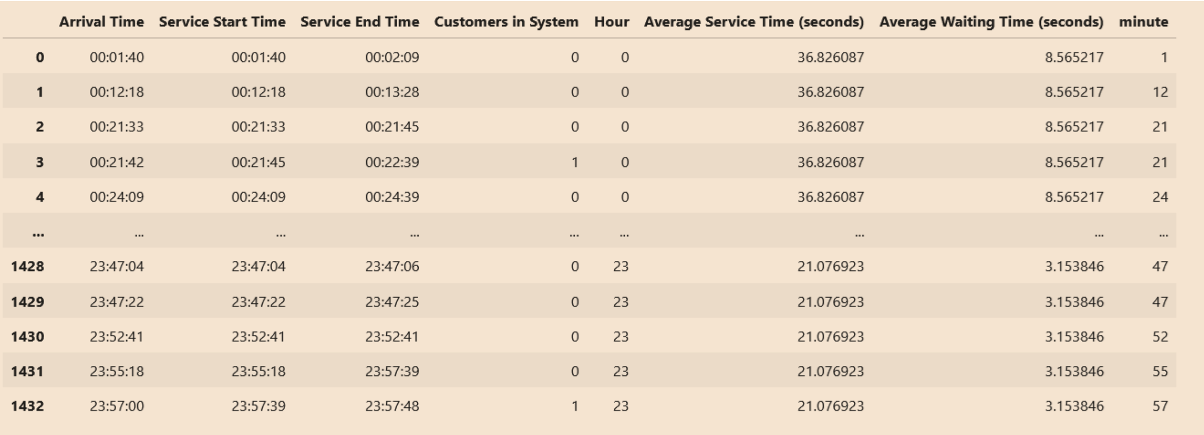
Implemented Random Forest Regression to predict customer waiting times based on historical queue data. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction. It is particularly effective for handling nonlinear relationships and interactions between features.

The input features used for the model include:

- Hour and minute of customer arrival
- Service duration
- Current queue length or number of customers waiting

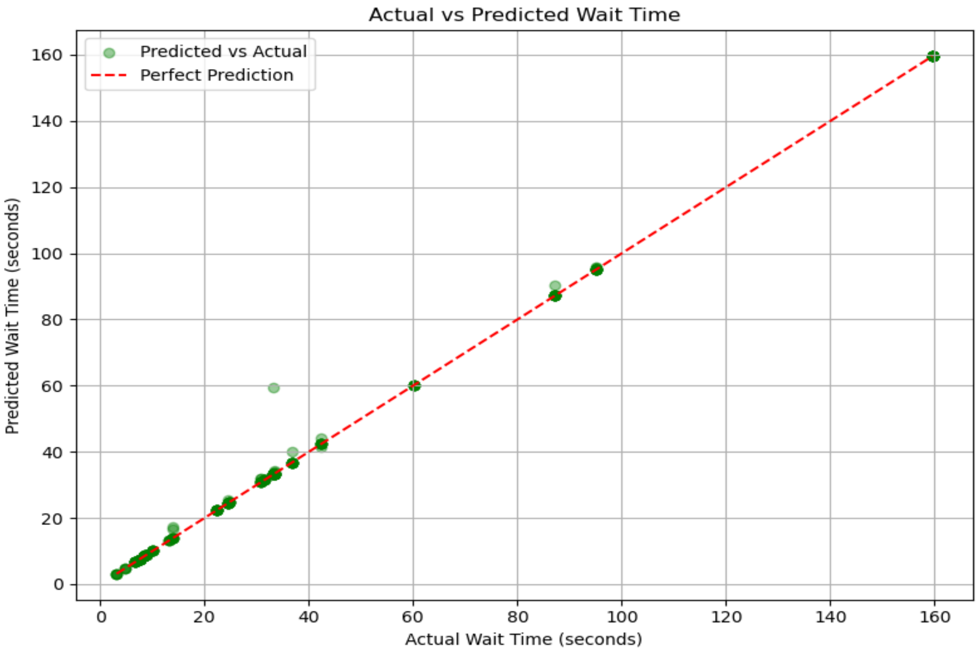
Performance is evaluated using Mean Absolute Error (MAE) and R² score on a held-out test set. Random Forest was selected due to its robustness, interpretability, and minimal preprocessing requirements.

Since obtaining real-world queue data from service systems is often challenging, this study utilizes simulated data generated through Discrete Event Simulation (DES) based on a call centre scenario. The simulated call centre operates 24/7, and the dataset represents a single day of operations, processing a total of 1,433 call arrivals. The arrival rate is non-stationary, with noticeable peaks during the midday hours, reflecting realistic fluctuations in customer demand.



	Arrival Time	Service Start Time	Service End Time	Customers in System	Hour	Average Service Time (seconds)	Average Waiting Time (seconds)	minute
0	00:01:40	00:01:40	00:02:09	0	0	36.826087	8.565217	1
1	00:12:18	00:12:18	00:13:28	0	0	36.826087	8.565217	12
2	00:21:33	00:21:33	00:21:45	0	0	36.826087	8.565217	21
3	00:21:42	00:21:45	00:22:39	1	0	36.826087	8.565217	21
4	00:24:09	00:24:09	00:24:39	0	0	36.826087	8.565217	24
...	...	...	...	...	...	...	...	...
1428	23:47:04	23:47:04	23:47:06	0	23	21.076923	3.153846	47
1429	23:47:22	23:47:22	23:47:25	0	23	21.076923	3.153846	47
1430	23:52:41	23:52:41	23:52:41	0	23	21.076923	3.153846	52
1431	23:55:18	23:55:18	23:57:39	0	23	21.076923	3.153846	55
1432	23:57:00	23:57:39	23:57:48	1	23	21.076923	3.153846	57

Mean Absolute Error (MAE) of the Random Forest Regression model is obtained as 0.17 seconds, that means on average, the model's predicted waiting times deviate from the actual waiting times by less than a second. This is a very low error, suggesting the model is highly accurate in its predictions. The model achieved an R² score of 0.96, indicating that it explains 96% of the variance in waiting times. This high level of accuracy confirms that the model is effective in capturing the underlying patterns in the queue system and can reliably predict customer wait durations.



In the above figure, each point represents a prediction. If a point lies exactly on the red dashed line, the prediction is perfect. If it's above the line, the model overestimated the wait time. If it's below the line, the model underestimated the wait time. The red line indicates the perfect prediction the closer the green dots are to this line, the better the model’s performance.

Here most points are very close to the line, indicating high accuracy. There's low spread, meaning low error/variance in predictions.

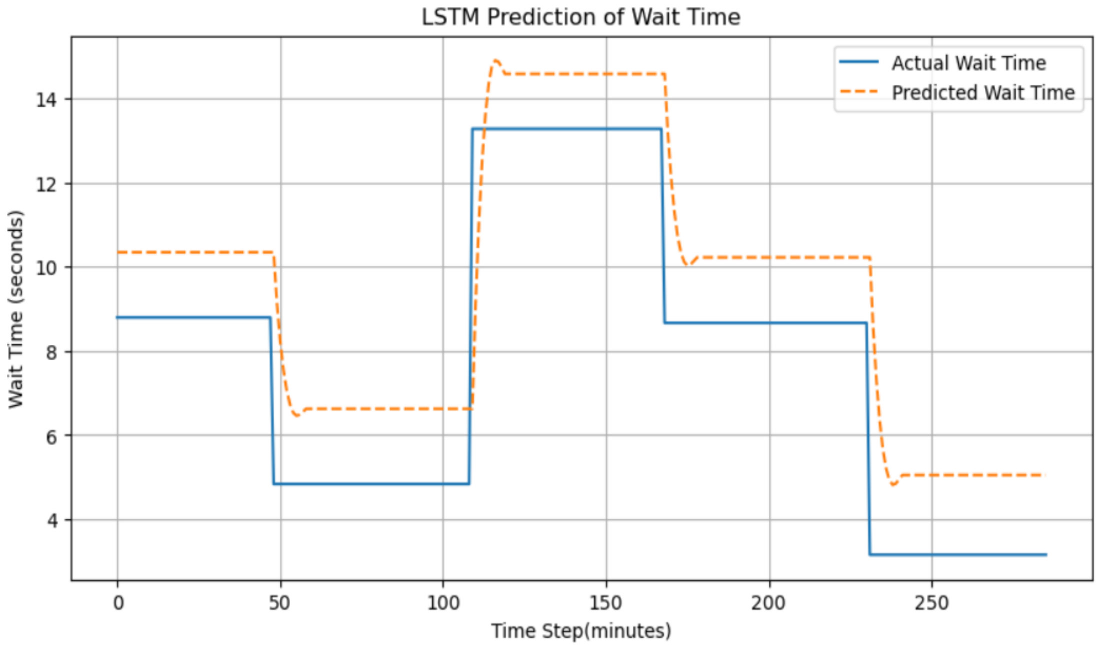
Forecasting Arrival Patterns with LSTM

LSTM networks are a type of recurrent neural network (RNN) well-suited for time series problems due to their ability to learn long-term dependencies in sequential data.

This model takes as input a sequence of past arrival counts (e.g., per 1-minute interval) and forecasts the expected arrivals for future time intervals.

The LSTM model achieved a Mean Absolute Error (MAE) of 1.76 seconds and an R² score of 0.7113, indicating that it explains over 71% of the variation in actual wait times. The model performs well in tracking time-dependent patterns, with small average errors and acceptable deviation.

In the figure given below, the blue line represents the actual wait times observed and the orange dashed line represents the LSTM model’s predictions. The model captures the overall shape and pattern of the wait times quite well and provides reasonably accurate forecasts. It performs well in identifying patterns but it reacts a bit slowly to sudden changes and smooths out sharp shifts. This is common in models like LSTM and can be improved by adjusting the model.



7. Business Modelling

QueueML is designed as a scalable and affordable AI-based solution for small and medium-sized service businesses such as clinics, diagnostic labs, local banks, call centres, and help desks. These businesses often face challenges such as long queues, unpredictable wait times, and inefficient staffing. QueueML addresses these issues by providing a machine-learning-driven wait time prediction tool and a real-time dashboard that helps service providers manage customer flow efficiently.

The product will be marketed through digital channels and partnerships with regional IT service vendors. Monetization will be achieved through a freemium pricing model with optional paid tiers offering additional features, as well as white-label licensing and advertising for the free version.

To build and operate QueueML, the core resources required include historical queue data, machine learning expertise, software developers, and cloud-based infrastructure. Key operational costs will include cloud server hosting, customer support, marketing, and ongoing software development.

Components	**Implementation**
Target Customers	Small clinics, diagnostic labs, local banks, call centres, customer service desks
Problems to be Solved	Long queues, unpredictable wait times, staffing inefficiencies
Core Product	ML-based wait time prediction system with a dashboard interface
Customer Acquisition Channels	Online marketing, partnerships with IT service providers, referrals

Revenue Streams	Subscription fees (freemium and tiered plans), white-label licensing, advertising
Key Resources	Queue data, ML algorithms, developers, cloud infrastructure
Cost Structure	Cloud hosting, salaries, support team, marketing

Monetization Ideas

To ensure financial sustainability and scalability, the following monetization strategies are proposed for QueueML:

1. Freemium Model:
   - Basic features (e.g., queue tracking, CSV upload, and single-location access) are offered for free.
   - Advanced features (e.g., real-time alerts, multi-branch dashboards, and forecast analytics) are part of a paid tier.
2. Tiered Subscription Plans:
   - Starter Plan: ₹500/month – for solo practitioners or small clinics.
   - Growth Plan: ₹1,000/month – for labs or multi-service centres.
   - Enterprise Plan: Custom pricing – for large chains with API integration needs and analytics support.
3. White Labelling and Licensing:
   - Allow regional health-tech companies or IT service firms to license and rebrand the product for local use.
4. Advertising and Sponsorship:
   - Free version users may see non-intrusive ads relevant to their industry.
   - Sponsored features can be integrated for partners (e.g., local diagnostic labs, service vendors).

8. Financial Modelling

Market Identification

- Market: Small & medium clinics, diagnostic labs, customer service centres (India)
- Approx. 100,000+ clinics and service centres in India.
- Target: Capture 1% (1,000 users) in 1 year.

Metric	**Value**
Monthly subscription	₹500 (Starter Plan)
Initial customers	30
Monthly growth rate	15%
Monthly operating cost	₹2,000 (hosting, support, etc.)

Basic Revenue Equation:
Let:

- x = number of customers in a given month
- y = monthly revenue

Then:

$$y = 500x - 2000$$

9. Conclusion

QueueML is an ML based solution designed to help small and medium service businesses manage queues more effectively. With a simple interface, smart machine learning features, and a flexible business model, QueueML can improve customer satisfaction, help staff work more efficiently, and make queue management easier for service providers.