Delivery Eligibility Using Geospatial Data

Determining Same-Day Delivery in Massachusetts, USA

Prabin Nepal  
 Data Foundation 6150-01  
 Wentworth Institute of Technology   
 Boston MA USA  
 [nepalp@wit.com](mailto:nepalp@wit.com)

ABSTRACT

This project investigates same-day delivery eligibility for addresses within Massachusetts, U.S., leveraging geospatial data from OpenAddresses and MassGIS. The study quantifies the relationship between distance, geographic constraints, and delivery feasibility by analyzing delivery zones defined through radius-based boundaries. Predictive regression models and efficient geocoding and visualization techniques are employed to ensure scalability and clarity.

KEYWORDS

Geospatial Analysis, Delivery Zones, Same-Day Shipping, Geocoding, Massachusetts

1 Introduction

Same-day delivery has emerged as a key differentiator for e-commerce platforms, requiring precise geospatial analysis to determine delivery feasibility. In Massachusetts, geographic boundaries such as rivers, highways, and urban centers add complexity to defining and analyzing delivery zones. This project explores how delivery zones can be mapped and analyzed to ensure efficient decision-making for same-day eligibility. The objectives include:

Defining delivery zones for Massachusetts using radius based methods.

Evaluating the impact of distance and geographic restrictions on delivery feasibility.

Developing efficient, scalable methods for analyzing multiple addresses.

By focusing on Massachusetts, this project addresses a specific U.S. context, providing actionable insights for delivery network optimization.

This project addresses a specific U.S. context by focusing on Massachusetts, providing actionable insights for delivery network optimization.

2 Data

The datasets used for this project were obtained from

credible sources and supplemented with manually generated data to fulfill the specific project objectives.

**MassGIS (Massachusetts Geographic Information System)**: The primary geospatial data source for this project was MassGIS, the official state agency providing geographic data for Massachusetts. MassGIS data is reliable and widely used for official and academic purposes. The dataset used included shapefiles defining the boundaries of towns and cities across the state.

**OpenAddresses:** Address data was obtained from OpenAddresses, a free, open-source repository that aggregates address points worldwide. The Massachusetts dataset included latitude, longitude, street names, and postal codes for various locations.

**Manually Generated Data:** To train and test the machine learning model, random points outside the delivery radius were generated programmatically using Python. The Haversine formula was used to calculate distances, ensuring that these points fell outside the 20-mile delivery zone. Each point was labeled as not eligible (label = 0) for delivery.

2.1 Data Cleaning and Preprocessing:

Geospatial data from MassGIS was reprojected to EPSG:4326 (WGS84 format) to ensure compatibility with address datasets. Points intersecting the 20-mile radius were extracted and labeled as eligible. Address data from OpenAddresses was standardized, and relevant columns were renamed (e.g., LON to longitude). Distances were calculated for all data points using the Haversine formula, ensuring accurate spatial measurements.

**Data Combination**:  
The MassGIS shapefile was used to define the delivery zone boundaries, and the OpenAddresses dataset provided real-world addresses for analysis. These datasets were combined by overlaying the delivery zone boundary on the address points to identify intersecting features. Additionally, manually generated points outside the radius were concatenated with the labeled data to create a comprehensive dataset for training and testing.

New Categories Created:

An eligibility label (label) was created for the model, based on whether a point was within or outside the 20-mile radius:

1: Eligible for same-day delivery.

0: Not eligible for same-day delivery.

The distance to the business (distance\_to\_business) was computed for each point to provide an additional feature for the model.

3 Methodology

This project combines geospatial analysis and machine learning to determine same-day delivery eligibility for addresses across Massachusetts. A 20-mile radius around the business location defines the delivery zone, and a Random Forest classifier predicts eligibility based on geographic features and distance calculations.

3.1 Geospatial Analysis

Geospatial methods were applied to define and visualize the delivery zone. A buffer radius of 20 miles was created around the business location, and addresses falling within this zone were identified and labeled.

Key Steps:

1. Data Projection:  
   All datasets were reprojected to a common coordinate system (EPSG:4326) to ensure compatibility between geographic layers.
2. Buffer Creation:  
   A 20-mile delivery radius was generated using geospatial tools.
3. Intersection Filtering:  
   Address points within the delivery zone were identified through spatial intersection with the buffer.
4. Visualization:  
   The delivery zone and eligible points were plotted using Matplotlib for clarity and presentation.

Tools Used:

* geopandas for buffer creation and intersection filtering.
* shapely.geometry for geometric operations.
* matplotlib for visualizing the delivery zone and eligible points.

3.2 Haversine Distance Calculation

The Haversine formula was used to calculate the great-circle distance between each address and the business location. This metric quantifies the proximity of addresses to the delivery zone.

Formula:

A math equations and formulas

Description automatically generated with medium confidence

Feature Engineering:

* Calculated distances for all points within and outside the delivery zone.
* Added distance\_to\_business as a feature to the dataset.
* **Validation**:  
  Distances were verified to match geospatially accurate measurements.

Tools Used:

* Python’s math library for trigonometric calculations.
* pandas for feature addition to the data

**3.3 Random Forest Classifier**

The Random Forest classifier was chosen for its robustness and ability to handle non-linear relationships. It predicts same-day delivery eligibility (label: 1 = eligible, 0 = not eligible) based on features like longitude, latitude, and distance\_to\_business.

* The datasets used for this project were obtained from credible sources and supplemented with manually generated data to fulfill the specific project objectives.Points inside and outside the delivery radius were combined into a single dataset.
* Real-world addresses from sampled\_addresses.csv were processed to add Haversine distances.
* The dataset was split into training (80%) and testing (20%) sets.

1. Model Training:
   * A Random Forest model with 100 decision trees (n\_estimators) was trained.
2. Model Evaluation:
   * Achieved 100% accuracy on the test set.
   * Precision, recall, and F1-scores were all perfect, demonstrating the model’s effectiveness.

**Feature Importance:**

* The model identified distance\_to\_business as the most critical predictor, followed by longitude and latitude.

**Tools Used:**

* sklearn.ensemble.RandomForestClassifier for model training.
* sklearn.metrics for evaluating accuracy and generating the classification report.
* matplotlib for visualizing feature importance.

**3.4 Testing with Real-World Data**

The model was tested with real-world addresses from across Massachusetts. These addresses were manually sampled and processed to simulate actual delivery requests.

**Data Preparation:**

* Real-world addresses were stored in sampled\_addresses.csv.
* Haversine distances were calculated for each address to serve as input for the model.
* Prediction:
* The trained model classified each address as eligible or not.
* Results were visualized, with eligible points plotted alongside the delivery zone and intersecting features.

**Tools Used:**

* pandas for managing address data.
* geopandas for geospatial processing.
* matplotlib for plotting predicted eligible points.

4 Results

The results of this study demonstrate the efficacy of integrating geospatial analysis and machine learning for determining same-day delivery eligibility. The Random Forest classifier achieved a perfect accuracy of 100% on the test dataset, with precision, recall, and F1-scores of 1.00 for both eligible (label = 1) and non-eligible (label = 0) points. These metrics highlight the model's exceptional ability to distinguish between addresses inside and outside the delivery zone. The geospatial visualization of the 20-mile radius delivery zone further validates these findings, as all predicted eligible points align precisely within the defined zone. This underscores the robustness of the methodology, where distance\_to\_business played a critical role in ensuring accurate predictions. Overall, the results confirm that this approach is highly effective for assessing delivery feasibility, providing a scalable and precise solution for delivery network optimization.

**4.1 Delivery Zone Definition**

The first visualization (Figure 1) illustrates the 20-mile delivery zone around the business location. The red boundary represents the buffer radius defining the delivery zone, while the blue polygons indicate intersecting features from the Massachusetts shapefile. The green star marks the business location. This plot clearly defines the spatial extent of eligible areas based on the geospatial intersection methodology.

This visualization confirms the precise definition of the delivery zone using geospatial analysis. By overlaying the intersecting features, it provides a clear understanding of the eligible geographic area for same-day delivery.

A screenshot of a map

Description automatically generated

***Figure 1****: The visualization demonstrates the effectiveness of using geospatial buffers to define delivery zones, ensuring that all intersecting areas within the radius are considered.*

**4.2 Prediction on Real-World Data**

The second visualization (Figure 2) builds on the delivery zone by integrating the model's predictions for real-world addresses. The orange points represent addresses predicted to be eligible for same-day delivery, overlaid on the existing delivery zone.

This plot showcases the model's ability to classify real-world addresses based on the trained Random Forest classifier.

A screenshot of a map

Description automatically generated

**Figure 2:** *The visualization confirms that all eligible points lie within the delivery zone, validating the model's accuracy in identifying delivery eligibility.*

The predicted points align closely with the defined delivery zone, with no significant outliers. This indicates that the model successfully generalizes the delivery eligibility logic to new, unseen data.

5 Discussion

While the project achieved excellent results with a perfect accuracy of 100% and demonstrated the utility of integrating geospatial analysis and machine learning, it is important to acknowledge potential limitations and areas for improvement. Addressing these issues will enhance the model's applicability to real-world scenarios and improve its robustness.

5.1 Limitations

* The model's reliance on a fixed 20-mile radius for defining the delivery zone does not account for dynamic factors such as traffic conditions, road closures, or weather constraints. This static approach limits the model's ability to adapt to real-world variability.
* The perfect accuracy of the model on the test data suggests a risk of overfitting. This could mean the model is overly tailored to the training data and may struggle to generalize to new, unseen datasets, particularly in scenarios with slight variations in geographic or operational features.
* The model currently uses only longitude, latitude, and distance\_to\_business as features. While effective for this study, the exclusion of operational metrics such as order cutoff times, traffic conditions, and geographic constraints may limit its real-world applicability.
* The model was tested on a small dataset of sampled addresses, which may not fully represent the geographic diversity or operational complexities of Massachusetts. Broader testing is required to validate the model's performance across different regions and scenarios.

5.2 Future Work

* To mitigate overfitting, future iterations of the model could increase the diversity and size of the training dataset by incorporating addresses from varied geographic contexts. Apply regularization techniques within the Random Forest classifier to prevent overfitting and perform cross-validation during training to ensure the model generalizes well across different data splits.
* Incorporating real-time traffic data and geographic barriers into the model would enable the creation of dynamic delivery zones based on travel time rather than a fixed radius. This approach would better reflect real-world delivery feasibility.
* Including additional metrics such as order cutoff times, package volumes, delivery priorities, and real-time weather data could improve the model's decision-making capabilities and make predictions more relevant to operational needs.
* Scaling the model to test addresses across different states and business locations would validate its robustness and scalability. Using datasets with diverse geographic and demographic characteristics would further ensure generalization.
* Future work could involve integrating the model into a live system that leverages GPS data, traffic updates, and order information to provide real-time delivery eligibility predictions.

6 Conclusion

This project demonstrated the successful integration of geospatial analysis and machine learning to predict same-day delivery eligibility for addresses in Massachusetts. By utilizing a 20-mile radius to define the delivery zone and applying a Random Forest classifier trained on features such as geographic coordinates and Haversine distance, the model achieved a perfect accuracy of 100%. The high precision, recall, and F1-scores validate the effectiveness of the methodology in distinguishing between eligible and non-eligible delivery points. The visualizations further reinforced the robustness of the approach, showing precise alignment between the model’s predictions and the defined delivery zone.

The results have significant real-world implications for optimizing delivery networks, enabling businesses to make informed decisions about delivery feasibility. The methodology is scalable, adaptable, and provides a foundation for incorporating additional real-world metrics, such as traffic and operational constraints, in future iterations. Despite limitations, such as potential overfitting and the static definition of delivery zones, this project establishes a strong starting point for further refinement and expansion in geospatial delivery optimization systems.

ACKNOWLEDGMENTS

This project was made possible by the resources provided by MassGIS and OpenAddresses for geospatial and address data. I would like to thank Wentworth Institute of Technology for supporting this research and for providing access to computational resources that enabled the analysis and visualization. Special thanks to faculty and peers for their guidance and feedback throughout the development of this project.

REFERENCES

OpenAddresses - U.S. Northeast. 2017. *OpenAddresses - U.S. Northeast*. Kaggle. Published on July 27, 2017. Accessed on December 8, 2024. URL: <https://www.kaggle.com/datasets/openaddresses/openaddresses-us-northeast?select=ma.csv>.Conference Name:ACM Woodstock conference

MassGIS. 2020. MassGIS Data: 2020 U.S. Census Towns. *Mass.gov*. Accessed December 8, 2024. Available at <https://www.mass.gov/info-details/massgis-data-2020-us-census-towns>.

Python Software Foundation. Python Programming Language. Accessed December 8, 2024. Available at <https://www.python.org>.

scikit-learn. RandomForestClassifier. *scikit*. Accessed December 8, 2024. Available at <https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.

Esri Community. 2021. Distance on a Sphere: The Haversine Formula. *Esri Community*. Published December 12, 2021. Accessed December 8, 2024. Available at <https://community.esri.com/t5/coordinate-reference-systems-blog/distance-on-a-sphere-the-haversine-formula/ba-p/902128#:~:text=For%20example%2C%20haversine(%CE%B8),longitude%20of%20the%20two%20points>.

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00