Trucking Market Research: Customer Segmentation and Targeting using Publicly Available Data

Data Science Institute
Columbia University
December 19, 2025

Ben Sullivan jbs2278@columbia.edu
Jared Donohue jjd2203@columbia.edu
Moacir P. de Sá Pereira mpd2149@columbia.edu
Jialin Wen jw4651@columbia.edu

Abstract

DrivePoints is an early-stage insurance technology company seeking to expand its customer base in the trucking industry. However, among the roughly 15 million registered trucks in the U.S., the company faces the challenge of efficiently identifying ideal clients for its insurance product. Leveraging a publicly available monthly census dataset from the U.S. Department of Transportation containing approximately 2 million trucking companies, we present a scalable lead generation tool. Our three-stage pipeline: (1) filters companies based on data completeness, validity, and recency, (2) predicts company fit using a logistic regression model trained on $n = 1,000$ annotated companies (AUC = 0.88), and (3) provides an interactive Streamlit dashboard for exploratory company filtering. Our approach demonstrates how publicly available government datasets can be transformed into actionable business intelligence for targeted customer acquisition.

1. Introduction

1.1 Background & Motivation

DrivePoints is an insurance technology company providing fleet management and insurance optimization that enables real-time tracking of trucks, promotes safe driving, and rewards both drivers and managers for safe behaviors ([1]).

As DrivePoints seeks to grow its customer base, it must decide which trucking companies are the most fitting for its product. With 14.89 million single-unit (2-axle, 6-tire or more) and combination trucks registered in 2023, representing 5% of all motor vehicles registered ([2]), it is not feasible to approach each company. Additionally, DrivePoints has nuanced criteria for the companies it wishes to target, such as not carrying hazardous material, making the identification of ideal companies a challenge.

1.2 Problem Statement

DrivePoints currently lacks a data-driven approach for identifying trucking companies that are well suited to its insurance product. While they have identified publicly available trucking census data, the data is missing values, has inconsistent reporting practices, and contains noisy fields, making it unclear whether such data can support reliable company-level targeting. The central problem addressed in this work is whether the federal transportation data can be transformed into a trustworthy, interpretable, and scalable targeting system that ranks trucking companies by their likelihood of being a good fit for DrivePoints insurance.

1.3 Project Overview

To address this problem, we develop an end-to-end pipeline for data quality assessment, interpretable statistical modeling, and operational deployment through an interactive dashboard application. Specifically, this project:

1. Assesses the completeness, consistency, and temporal stability of the FMCSA Motor Carrier Census and quantifies its suitability for downstream modeling through explicit data quality metrics.
2. Constructs a probabilistic company fit model using annotated examples to estimate the likelihood that a trucking company aligns with DrivePoints' underwriting and operational criteria.
3. Compares deterministic rule-based filters, statistical models, and large language model–assisted validation to understand the tradeoffs between accuracy, interpretability, and scalability.
4. Delivers an interactive dashboard that enables stakeholders to explore, filter, and rank trucking companies based on model scores and key operational attributes.

2. Methods

2.1 Data

2.1.1 Data Sources

The primary data source is provided by the Federal Motor Carrier Safety Administration (FMCSA) ([3]). The FMCSA is a part of the US Department of Transportation (USDOT) aiming to reduce crashes and injuries involving large trucks and buses. They publish a monthly Motor Carrier Census (MCC) as part of the Safety Management System (SMS). This forms the center of our dataset and the one that we aim to analyze monthly.

The dataset contains 42 columns, including the USDOT Number, company names, addresses, contacts, telephone and fax numbers, e-mail, HazMat flag, passenger carrier flag, number of power units, number of drivers, mileage, mileage year, operation, and classification registration information. The file is comma delimited with one carrier per row. Each monthly installment is called a "version," and the most recent version (110, released December 13, 2025) has 2.08m rows.

In addition to the FMCSA MCC dataset, we identified three supplemental data sources to improve DrivePoints business targeting.

1. **US DOT Insurance History ([4]).** This dataset contains information on a carrier's, broker's, or freight forwarder's previous insurance policy(ies), including the DOT number for each business, an identifier that allows for joining with the MCC table. The dataset describes what kind of insurance a company had, how they canceled, and what the policy was. The data pertain only to previous policies, so it does not include any company's current insurance. The dataset contains 7.1m rows and 18 columns. Data updated 9/24/2025.

2. **Fatality Analysis Reporting System (FARS) ([5]) and Crash Report Sampling System (CRSS) ([6]).** Both the FARS and CRSS datasets are released annually, and for the purposes of our analysis, we are looking at data from 2020 to 2023. The FARS data lists every fatal crash in the US. It includes details of the crash, such as the vehicle, number of victims, driver violations (including license suspensions), circumstances of the crash that can be used to infer fault, and DOT number (when applicable). The combined FARS dataset from the last four years contains records from 235,438 crashes, of which there are 13,712 records with DOT numbers available. Meanwhile, the CRSS dataset contains a representative sample of all vehicle crashes, so it is important to note that many crashes go unreported in this data. It generally contains the same fields as the FARS dataset, and the same inferences can be made. Over the last four years of available data, there are 372,720 records, of which 6,971 have usable DOT numbers reported.

3. **Cargo Carried (FMCSA QCMobile API)** ([7]). Cargo type is a critical feature for DrivePoints' underwriting and targeting decisions, as it directly indicates whether a motor carrier transports goods aligned with DrivePoints' insurance products (e.g., building materi-

als versus passengers or household goods). FMCSA exposes cargo information through the SAFER Company Snapshot: `https://safer.fmcsa.dot.gov/CompanySnapshot.aspx`. It provides a concise electronic record of a carrier's identification, size, commodity information, and safety history, including a categorical `cargo_carried` field. Cargo carried is categorized into 29 predetermined default types, with a textbox option for carriers to fill in. This provided many "other" types of cargo that needed to be further categorized. Ultimately, we added 7 new categories to the pre-existing categories that we then used in our model and classified approximately 75,000 unique tokens into these 36 total categories.

Initially, when API access was unavailable due to the government shutdown, we extracted the cargo carried field for a sample of records by scraping the SAFER Company Snapshot website using a Python script ([8]). However, this would not scale, so once the FMCSA QCMobile API was restored we transitioned to a fully programmatic approach, querying the `/carriers/dotNumber/cargo-carried` endpoint to retrieve cargo classifications in structured JSON format without the need for scraping. This substantially improved scalability, reproducibility, and data reliability, and enabled systematic enrichment of the Motor Carrier Census using USDOT numbers as join keys.

2.1.2 Data Preprocessing

A few preprocessing steps make the FMCSA MCC dataset ready for efficient analysis. These include: (1) converting Motor Carrier Census data from CSV to parquet, (2) down-casing column names, (3) converting `add_date` and `mcs150_date` fields to dates, and (4) converting zeroes to `null`s in recent mileage columns. [1]

[1] When the code for the source of the vehicular miles traveled (`vmt_source_id`) is null, `recent_mileage` is not null. Instead, it is given as 0, as is the year for the recent mileage (`recent_mileage_year`). We believe encoding these mileage and year values as null instead of 0 is appropriate.

2.2 Exploratory Data Analysis

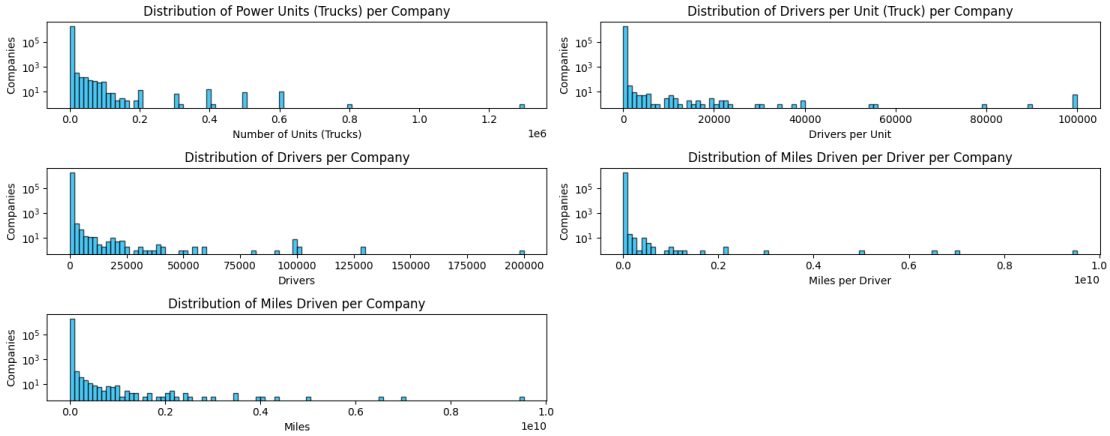

Figure 1: Distribution of Key Numerical Fields (using December 2025 data)

Our exploratory data analysis evaluated the latest version of the census as well the dataset over time. In general, we can see that the numerical fields have a long tail with some unlikely outliers, particularly in ratios we calculate, such as the number of miles driven per driver per company.

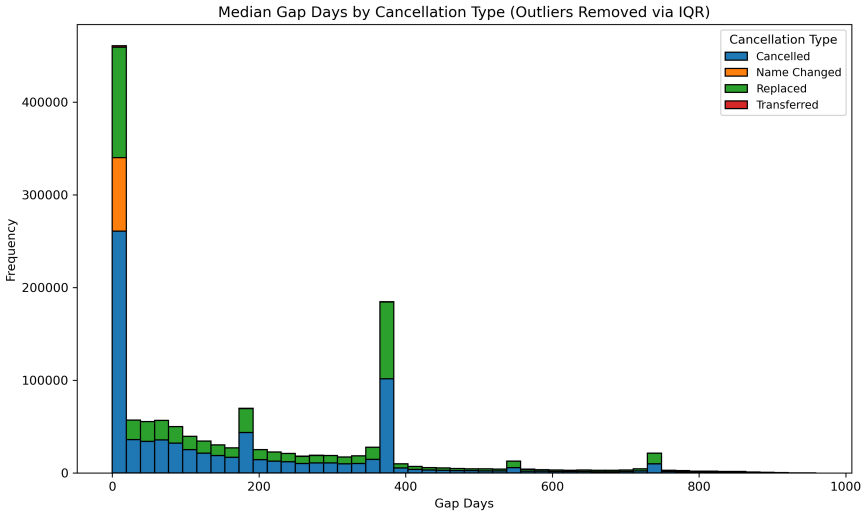

Figure 2: Median Time Between Insurance Filings

2.2.1 Missing Data Analysis

Although the most `null` columns are fax numbers, "doing business as" (DBA) names, and email addresses, the following columns have implications for the utility of the census data and are analyzed with more scrutiny. These numbers are for the October report (version 108) ([9]).

Column	Description	Null Fraction
vmt_source_id	VMT source	0.59
mcs150_mileage_year	Year of MCS-150 mileage report	0.38
mcs150_mileage	VMT reported in MCS-150	0.34
mcs150_date	Last MCS-150 date	0.08
nbr_power_unit	Number of power units (fleet size)	0.03

VMT refers to "vehicular miles traveled," and the census provides for three sources of that number: a company's MCS-150 filing, an audit, or an investigation. The MCS-150 or Motor Carrier Identification report is a form companies are required to file to register for a DOT number. The report must be refiled biannually, with the reporting deadline signaled by the final two digits of the DOT, so that every month a portion of the companies have to report over the source of the two years. Companies must also file a new MCS-150 if there is a change in the information. However, an analysis of the relationship between the MCS-150 columns and VMT source indicates that nearly all of the null values for the MCS-150 data are explained by nulls in the VMT source. Hence, if there is no recent VMT source, there is likely no MCS-150 data.

2.2.2 Number of Drivers & Mileage Data Analysis

The driver-to-truck ratio indicates that the large majority of companies have a few drivers per power unit. The 75% quartile was 1, while the mean was 2.2. It is probably unlikely that any companies with thousands of drivers per truck are reporting their data correctly, suggesting a ratio greater than 3:1 or so might induce skepticism of the data for that company.

Similarly, the mileage ratios per truck and per driver also show extremely unlikely values. When we apply thresholds to the outliers (20 drivers per truck and 200k miles per truck or driver), we see that the resulting outliers make up less than 1% of the data.

2.2.3 Historical Analysis

For historical analysis, we downloaded the last 30 editions of the data, stretching back over two years (and so collecting an entire cycle of MCS-150 reports). Our first comparison was to see the stability of null fractions over time. The same columns lead the pack as in the single-table analysis, and by effectively the same fractions. Our analysis of the distributions of the null values indicate that about the same amount of data is missing from version to version. Sparse data remain sparse, and dense data remain dense.

Next for us was to determine if a single company (identified by a unique DOT number) varied widely in its reporting, particularly around these mileage, fleet, and driver-related data points. We see that the recent values are directly tied to whether or not a VMT source was reported, while the MCS-150 reporting is also tied together on a per-company basis. The companies that report

MCS-150s always report them. Finally, as we already knew, nearly all companies always report both fleet sizes and driver counts.

Column	Intermittently Null	Always Null	Never Null
vmt_source_id	0.07	0.58	0.35
recent_mileage_year	0.07	0.58	0.35
recent_mileage	0.07	0.58	0.35
mcs150_mileage_year	0.02	0.38	0.60
mcs150_mileage	0.025	0.34	0.63
nbr_power_unit	0.001	0.04	0.96
driver_total	0.0001	0.001	0.9985

2.3 Analytic Methods

2.3.1 Data Quality Scoring (DQS)

To assess the quality of the FMCSA Motor Carrier Census data, we developed a composite Data Quality Score (DQS) metric that quantifies dataset-level and record-level usability for market targeting.

Each data record is evaluated along three dimensions to assess the data quality: completeness, validity, and timeliness. Completeness measures the proportion of required fields populated with non-null, non-placeholder values across key identifiers and operational attributes (e.g., legal name, contact information, MCS-150 filing date). Validity assesses whether provided values are structurally and semantically plausible, including format checks for phone numbers, ZIP codes, and email addresses, as well as cross-field consistency checks (e.g., city–state–ZIP coherence). Timeliness captures whether operational filings reflect recent carrier activity, with particular emphasis on the recency of MCS-150 updates.

The DQS is computed as a weighted linear combination of these dimensions, with weights selected to reflect their relative importance for reliable targeting based on exploratory analysis and DrivePoints business context. Completeness and validity are weighted more heavily than timeliness, as missing or malformed fields fundamentally limit usability even for recently active firms.

$$\text{DQS} \;=\; 0.4 \times \text{Completeness} \;+\; 0.4 \times \text{Validity} \;+\; 0.2 \times \text{Timeliness}.$$

Scores are calculated for all records in the dataset, enabling both population-level quality assessment and per-carrier filtering. Records with DQS exceeding a predefined threshold (e.g., 0.7) are retained for downstream total addressable market (TAM) estimation and model-based targeting, while lower-quality records are excluded to reduce noise and wasted outreach effort. This approach follows established data quality frameworks while remaining tightly coupled to operational decision-making needs.

2.3.2 Rule-Based Filtering to estimate Total Addressable Market (TAM)

To identify the total addressable market (TAM) of trucking companies that meet DrivePoints' insurance targeting criteria, we initially implemented a rule-based filtering guided by DQS and mentor correspondence.

The filtering logic first applies a set of exclusion rules that disqualify companies inconsistent with DrivePoints' target profile. Firms are removed if they (1) operate primarily under *interstate* authority (`carrier_operation = "A"`), (2) handle hazardous materials (`hm_flag = TRUE`), (3) are located in states and territories in which DrivePoints does not wish to grow (New Jersey, New York, Alaska, Hawaii, or Puerto Rico), (4) are government, school, or public entities, or (5) show inactive or incomplete operational data. These exclusion criteria directly correspond to the "BAD" category defined in the mentors' annotated prompt and email guidance.

Companies that pass the exclusion filter are then evaluated against positive indicators that align with DrivePoints' ideal customer profile: being *authorized for hire* or classified as *private carriers*, maintaining a fleet size between 1 and 50 trucks (with 3–20 as the operational sweet spot), operating in the five priority states (California, Texas, Arizona, Utah, and Nevada), and exhibiting recent operational activity (MCS-150 filings from 2023–2025 with plausible annual mileage and driver ratios). Firms meeting these operational and geographic criteria are considered part of the Total Addressable Market (TAM) for DrivePoints' products.

This rule-based approach ensures that eligibility decisions remain transparent, auditable, and scalable, while maintaining alignment with the business definitions initially validated through LLM experiments and human annotation.

2.4 LLM Experiments

We evaluated whether large language models (LLMs) could support (1) data quality assessment and (2) overall company fit scoring using DrivePoints-specific underwriting criteria. The goal was to determine whether LLMs could replace or meaningfully augment deterministic rules for large-scale customer targeting.

We conducted controlled experiments using Gemini 2.5 model variants (Flash Lite, Flash, and Pro), providing each model with structured FMCSA records, explicit data quality definitions, and DrivePoints-specific eligibility criteria ([8]). Model outputs were evaluated against independent human annotations using binary accuracy and balanced accuracy to account for class imbalance ([10]).

Across model variants, LLMs achieved binary accuracy between 72–76% on a sample of $n = 100$ records. However, this performance masked a pronounced classification bias. The models consistently over-labeled companies as `BAD`, achieving high recall for negative cases (approximately 88%) but very low recall for positive cases (approximately 26%). As a result, balanced accuracy remained low at 55–57%, indicating poor calibration for identifying high-quality companies suitable for targeting.

Despite this limitation, LLMs demonstrated strengths in validity assessment. They reliably flagged nuanced issues such as implausible mileage values, outdated MCS-150 records, and non-trucking entities—cases that are difficult to capture with static rule-based checks. These findings suggest that LLMs can add value as a secondary audit mechanism for data quality, but not as a primary tool for ranking or identifying acquisition targets.

Comparisons across Gemini model variants showed minimal differences in predictive performance despite substantial differences in cost and runtime. Given the lack of material accuracy gains from rule-based filtering, cost and latency considerations further weakened the case for LLM-based production scoring.

Overall, these experiments showed that while LLMs provide useful contextual reasoning for targeted validation and anomaly detection, their classification bias and limited balanced accuracy make them unsuitable as a standalone solution for large-scale customer acquisition. As a result, our production system relies on deterministic rules and interpretable statistical models, with LLMs reserved for future complementary quality audits where human review is already warranted.

Prompt	Model	Sample Size	Binary Accuracy	Balanced Accuracy
V1 (baseline)	gemini-2.5-flash-lite	$n = 100$	49.0% ± 9.8%	35.0%
V2 (mentor instructions)	gemini-2.5-flash-lite	$n = 100$	72.0% ± 8.8%	55.0%
V3 (revised mentor instructions)	gemini-2.5-flash-lite	$n = 100$	72.0% ± 8.8%	55.0%

Table 3: Prompt Version vs Accuracy for LLM classification of company fit. Balanced accuracy is computed using the scikit-learn definition to account for class imbalance.

Model	Sample Size	Binary Accuracy	Balanced Accuracy	Runtime	Cost
gemini-2.5-flash-lite	$n = 100$	72.0% ± 8.8%	55%	2 min (1.2 sec/record)	$0.02
gemini-2.5-flash	$n = 100$	75.0% ± 8.5%	55%	25 min (15 sec/record)	$0.04
gemini-2.5-pro	$n = 100$	76.0% ± 8.4%	57%	22 min (12 sec/record)	$0.23

Table 4: Gemini Model vs Accuracy. Accuracy and balanced accuracy show minimal performance gains relative to large differences in cost and runtime.

Model	100 Records	1,000 Records	10,000 Records	100,000 Records	Full Dataset (2.09M Records)
gemini-2.5-flash-lite	$0.02	$0.20	$2	$20	$418
gemini-2.5-flash	$0.04	$0.40	$4	$40	$836
gemini-2.5-pro	$0.23	$2.30	$23	$230	$4,807

Table 5: Estimated costs for running Gemini models on increasing dataset sizes, including the full 2.09M-record dataset (assuming linear scaling).

2.5 Statistical Model

To generate a stable and generalizable *Company Fit Score* for over two million FMCSA carriers while relying on a small manually labeled dataset, we designed a modeling pipeline centered on (1) robust feature engineering, (2) systematic control of feature dimensionality, and (3) reliable model selection via stratified cross-validation. The goal is to prevent overfitting to the limited labeled states while still producing reliable predictions across all U.S. states and carrier categories.

2.5.1 Logistic Regression Framework

We use logistic regression as the core predictive model. After feature preprocessing, each carrier record is represented by a set of numerical variables and encoded categorical features. Logistic regression models the conditional probability

$$\hat{p} = P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}},$$

where $\hat{p}$ represents the likelihood that a company is a "good fit" for DrivePoints. This probability is used directly as the continuous *Company Fit Score* to rank all 2.09M carriers.

Logistic regression is well suited for our setting because it handles high-dimensional sparse features, remains interpretable, and—when paired with appropriate encoding—generalizes well under small-sample conditions.

2.5.2 Feature Engineering

We incorporated numerical and categorical fields from the enhanced FMCSA dataset. The numerical features were standardized using `scikit-learn StandardScaler` ([11]) to ensure stable coefficient estimation. These included features like: number of power units, number of drivers, and mileage.

For categorical features, the low-cardinality fields are one-hot encoded while high-cardinality fields, such as carrier operation type and cargo carried category, use smoothed target encoding. This preprocessing establishes the foundation for controlling sparsity and preventing overfitting.

2.5.3 Feature Dimension Smoothing and Imbalance Control

A central challenge in our modeling pipeline is the combination of (1) extremely limited labeled data and (2) several high-cardinality categorical features. Direct one-hot encoding of all categories produces hundreds of sparse dummy variables, many of which correspond to categories that appear only once or not at all in the labeled subset. This leads to severe overfitting and highly unstable coefficient estimates. To address these issues, we evaluated a sequence of encoding and dimensionality-control strategies, each motivated by its ability to reduce sparsity, smooth rare-

category effects, or mitigate label imbalance.

(1) Full categorical detail. All categorical levels are retained without grouping. This maximizes information but produces a highly sparse design matrix, as many categories appear infrequently or are absent in the labeled subset. Such sparsity makes model coefficients extremely sensitive to rare categories and increases overfitting risk.

(2) Top-5 truncation. For each high-cardinality feature, only the five most frequent categories are kept; all remaining categories are merged into a single "Other" bucket. This substantially reduces dimensionality and forces the model to avoid relying on extremely rare categories, serving as a simple but effective form of structural regularization.

(3) Adaptive top-k. Instead of fixing $k = 5$, each feature selects its own k based on the frequency distribution of its categories. While more flexible, this approach can still preserve many low-frequency categories if the distribution is relatively flat. As a result, the model may continue to overfit rare categories that appear only once or twice in the labeled sample.

(4) One-Hot Encoding. For categorical variables with naturally low cardinality (e.g., carrier operation type), we apply standard one-hot encoding. Each category is represented as a binary indicator. This encoding preserves clear interpretability and works well when every category appears sufficiently often in the labeled dataset.

(5) Target Encoding (TE). For high-cardinality categorical variables, we apply smoothed target encoding, which replaces each category with the estimated conditional mean of the target variable. To prevent overfitting on categories with few labeled observations, the category-level mean is shrunk toward the global mean using a smoothing factor. The encoding for a category c is computed as

$$\mathrm{TE}(c) = \frac{n_c \cdot \bar{y}_c + s \cdot \mu}{n_c + s},$$

where n_c is the number of labeled samples belonging to category c, $\bar{y}_c$ is the category-specific label mean, μ is the global label mean, and s is a smoothing parameter. When n_c is large, the estimate approaches the category mean; when n_c is small, the estimate shrinks toward the global mean. We implement TE in a leakage-free manner using stratified K-fold splitting, generating out-of-fold encoded values for training and a full-sample encoding map for production inference.

In high-cardinality categorical features, many categories appear only a few times in the labeled dataset, making their empirical means extremely noisy. For a category c with only $n_c = 1$ or 2 labeled samples, the raw estimate $\bar{y}_c$ has a large variance:

$$\mathrm{Var}(\bar{y}_c) = \frac{p(1-p)}{n_c},$$

which diverges as $n_c \to 0$. Without smoothing, the model may assign disproportionately large positive or negative coefficients to such rare categories, effectively memorizing individual samples and leading to unstable decision boundaries.

Smoothing counteracts this by shrinking the noisy category mean $\bar{y}_c$ toward the global mean μ, yielding the stabilized estimator

$$\mathrm{TE}(c) = \frac{n_c \cdot \bar{y}_c + s \cdot \mu}{n_c + s}.$$

From a Bayesian perspective, this is equivalent to placing a prior with strength s on μ, resulting in a maximum a posteriori (MAP) estimate. When n_c is small, the prior dominates and TE approaches μ; when n_c is large, the data dominate and TE approaches $\bar{y}_c$. This adaptively reduces variance for rare categories while preserving information for common ones, producing numerically stable and generalizable encodings under limited labeled data.

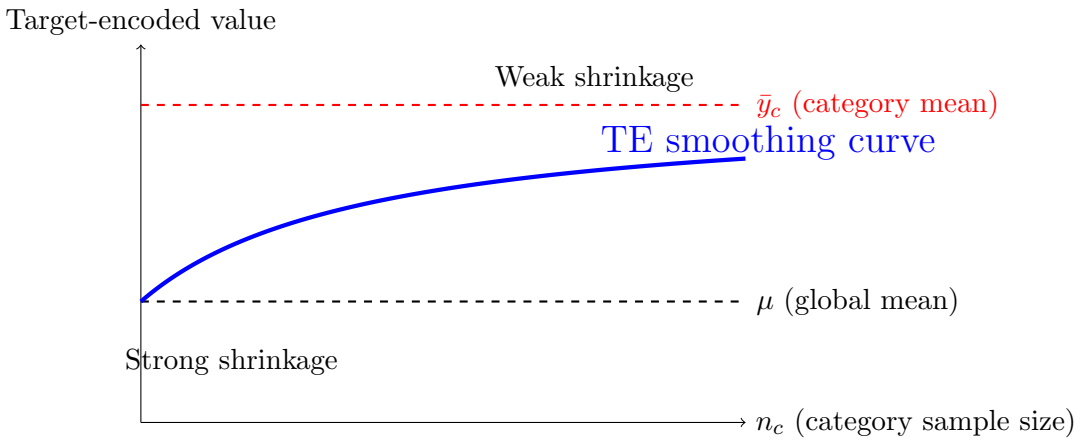

Figure 3: Illustration of smoothed target encoding. When the category sample size n_c is small, the encoding shrinks strongly toward the global mean. As n_c increases, the encoding converges to the category-specific mean.

To ensure stable model evaluation under label imbalance, we adopt stratified K-fold cross-validation. Stratification preserves the proportion of positive and negative labels in each fold, preventing folds from missing minority classes. This reduces variance in performance estimates and produces more trustworthy assessments of how well each encoding strategy generalizes.

Table 6: Comparison of feature dimension control and smoothing strategies.

Method	**Hold-out AUC**
TE + One-Hot + Cargo Information	0.870
TE + One-Hot	0.780
TE + One-Hot + stratified CV	0.770
Top-5 truncation	0.745
XGBoost baseline	0.730
Adaptive top-k	0.720

Many states and carrier types appear rarely or not at all in the labeled data. Without smoothing, TE will overfit tiny category counts. Smoothing shrinks category-level means toward the global mean when sample size is small, preventing extreme estimates and enabling full-coverage scoring across all carriers.

2.5.4 Tree-Based Ensemble Benchmarks

We benchmarked two commonly used tree-based ensemble models: Random Forests (RF) and XGBoost. We tested them using the same labeled dataset and encoded features. Both models underperformed logistic regression, scoring AUC between 0.70–0.73. Their limitations arise from the interaction between high-cardinality categorical variables and the small number of labeled samples:

- the encoded feature matrix remains high-dimensional relative to the labeled sample size;
- tree splits often operate on categories with very few observations, providing insufficient information to form stable partitions;
- many resulting leaf nodes contain only 1–2 labeled samples, leading to local memorization rather than generalizable patterns;
- rare states or carrier types cannot be reliably learned because tree ensembles do not incorporate smoothing mechanisms for low-frequency categories.

Although ensemble methods typically reduce variance through averaging or boosting, they cannot overcome the fundamental imbalance between dimensionality and labeled data availability in this setting. In contrast, logistic regression combined with smoothed target encoding provides stronger performance and more reliable generalization under limited-label, high-cardinality conditions.

2.5.5 Model Training and Selection Pipeline

We conduct grid search over regularization strength (C) using stratified 5-fold CV. L2 regularization consistently outperforms L1, which zeroes out many informative features under small-sample conditions.

After selecting the best hyperparameters, the final model is retrained on the full training set and applied to all 2.09M carriers:

$$\text{ml\_score} = P(Y = 1 \mid \text{features}).$$

This score forms the foundation for downstream ranking and dashboard presentation.

2.5.6 Model Interpretability

Logistic regression retains full interpretability through the coefficient sign and odds ratios. A positive coefficient sign increases the log-odds of being a good-fit carrier, and a a negative coefficient decreases it.

Exponentiating each coefficient (e^{β_i}) gives the multiplicative change in odds associated with a one-unit change in the feature.

These properties provide extremely transparent insights into the features which influence the Company Fit Score model output.

2.5.7 Model Evaluation

Model performance is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC), as our objective is to rank companies by predicted fit rather than make strict binary decisions. AUC is threshold-independent and robust to class imbalance, making it well suited for large-scale lead generation with limited labeled data.

Among all candidate specifications, the final model combining target encoding, one-hot encoding, and aggregated cargo information achieves the best ranking performance, with an AUC of 0.87 on the validation set. This model is therefore used for downstream company scoring and market sizing.

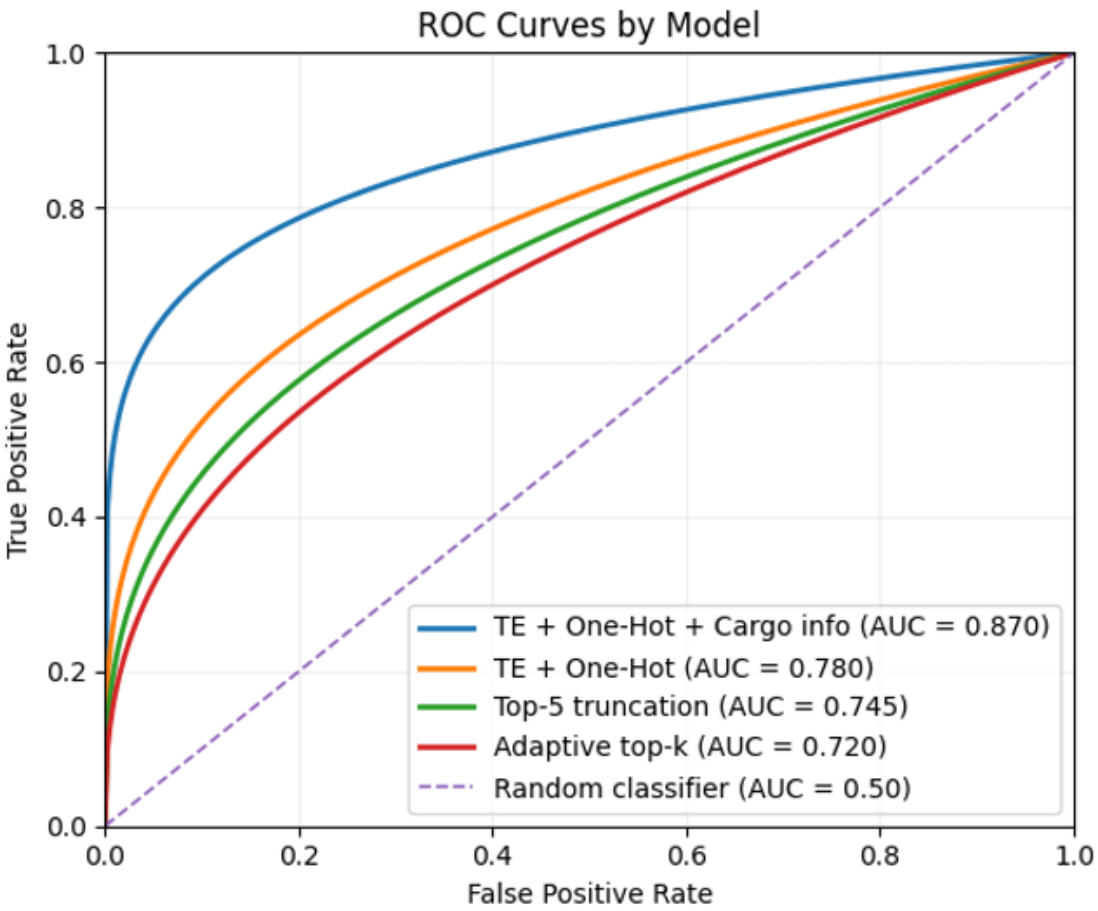

Figure 4: ROC Curve for Logistic Regression Model

Accordingly, when the objective is to identify as many suitable potential customers as possible, recall is prioritized over precision. By experimenting with different decision thresholds, we find that 0.21 represents the most aggressive operating point that still preserves reasonable classification quality. At this threshold, the model is able to retain approximately 90% of truly high-quality target companies without incurring a severe loss in accuracy.

As a result, we recommend using a Company Fit Score of 0.21 as the lower bound when exploring results in the dashboard. Including companies with scores below this level leads to diminishing returns, as further gains in recall become marginal while precision deteriorates rapidly. Therefore, a threshold of 0.21 provides a practical and interpretable balance between maximizing target coverage and maintaining lead quality in a real-world sales setting.

Table 7: Classification Performance at Company Fit Score Threshold = 0.21

Metric	Value
Decision Threshold	0.21
Recall	0.90
Precision	0.68
Accuracy	0.78
F1 Score	0.78

When the Company Fit Score threshold is increased to 0.8, the model adopts a highly selective targeting strategy. Under this setting, precision rises to 0.86, indicating that the majority of identified companies are indeed high-quality targets. However, this gain in precision comes at a substantial cost to recall, which drops to 0.44. In other words, more than half of the truly suitable companies are excluded at this threshold.

This result highlights a clear trade-off between lead quality and market coverage. While a threshold of 0.8 is effective for conservative targeting scenarios where precision is critical, it is less suitable when the objective is to comprehensively identify potential high-quality customers.

Table 8: Classification Performance at Company Fit Score Threshold = 0.8

Metric	Value
Decision Threshold	0.80
Precision	0.86
Recall	0.44
Accuracy	0.74
F1 Score	0.58

3. Results

3.1 Quality Assessment of Federal Trucking Census Data & Initial TAM

We calculate the Data Quality Score (DQS) of the USDOT FMCSA dataset to be 0.655 on a scale from 0 to 1, implying moderate dataset quality. Using DQS as a filter, our intial estimate of the Total Addressable Market (TAM) for DrivePoints was 550,884 companies. The full USDOT FMCSA dataset (2,091,643 records) was processed using a DQS threshold of 0.80, resulting in 1,288,458 high-quality records retained, from which 550,884 companies were identified as eligible targets within the Total Addressable Market (TAM). The DQS filtering pipeline processes the full dataset in roughly one minute end-to-end, and generates both detailed company-level outputs and state-level TAM summaries.

State	TAM Companies	TAM Trucks	Avg. DQS	Priority Share	Sweet Share
CA	78,947	188,331	0.866	1	0.187
TX	69,313	153,974	0.859	1	0.184
FL	49,494	105,755	0.857	0	0.162
GA	29,947	67,796	0.849	0	0.206
PA	24,047	67,050	0.855	0	0.278
MI	22,225	65,588	0.840	0	0.291
WI	20,873	58,217	0.850	0	0.278
MN	20,204	52,303	0.843	0	0.239
CO	17,835	50,930	0.848	0	0.269
WA	17,666	46,933	0.848	0	0.252

Table 9: Top 10 States by TAM after DQS and Rule-Based Filtering (Threshold = 0.80)

Note. **Priority Share** indicates whether the state is part of DrivePoints' prioritized operating regions (CA, TX, AZ, UT, NV). **Sweet Share** represents the proportion of carriers within the ideal fleet size range (3–20 trucks), highlighting concentration of the target customer segment.

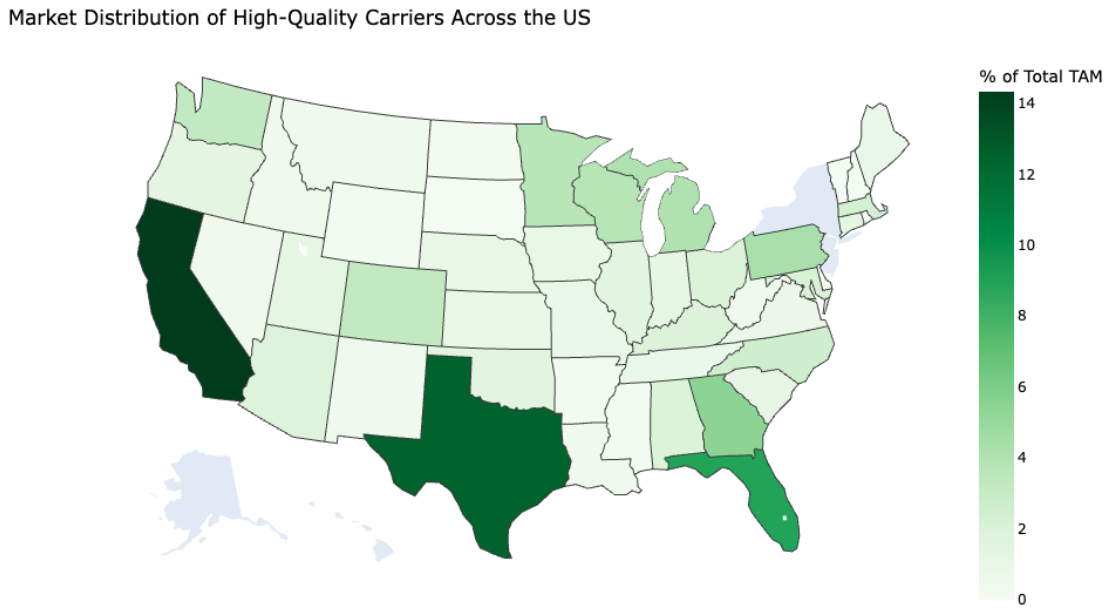

Figure 5: State-level TAM distribution after DQS and rule-based filtering

3.2 Statistical Model Predicting Best Companies to Target

As discussed in the Model Evaluation section, the statistical model outputs a continuous Company Fit Score, and a minimum operating threshold of 0.21 represents the most aggressive boundary for recall-oriented targeting. In practice, however, different threshold choices correspond to different business objectives and levels of lead selectivity.

To illustrate how threshold selection affects the size and composition of the target market, we report state-level summaries under two representative Company Fit Score cutoffs. Specifically, we present the top states by final Total Addressable Market (TAM) when the Company Fit Score threshold is set to 0.8, corresponding to a highly selective targeting strategy, and when it is set to 0.3, corresponding to a broader, recall-oriented strategy. In both cases, all companies are required to satisfy the data quality constraint of DQS ≥ 0.8. relevance.

Table 10: Top States by Final TAM with Company Fit Score ≥ 0.8 and DQS ≥ 0.8

State	TAM Companies	TAM Trucks	Avg DQS	Avg Fit
CA	56,893	874,224	0.869	0.922
FL	34,769	527,716	0.860	0.963
TX	23,770	398,648	0.881	0.884
GA	21,680	503,298	0.853	0.956
WI	13,003	139,056	0.855	0.962
CO	10,242	444,549	0.854	0.903
AL	6,910	121,947	0.847	0.928
MA	6,732	140,721	0.870	0.884
MI	6,165	165,946	0.863	0.888
PA	6,065	260,520	0.878	0.880

Table 11: Top States by Final TAM with Company Fit Score ≥ 0.3 and DQS ≥ 0.8

State	TAM Companies	TAM Trucks	Avg DQS	Avg Fit
CA	64,831	917,148	0.869	0.880
FL	55,800	602,485	0.871	0.760
TX	44,387	568,115	0.863	0.783
GA	32,269	565,000	0.864	0.782
NY	31,196	259,501	0.859	0.544
WI	18,771	196,021	0.862	0.808
PA	17,721	522,519	0.860	0.719
MI	14,434	217,350	0.844	0.752
CO	13,027	503,905	0.850	0.847
MN	12,982	178,841	0.847	0.705

3.3 Interactive Marketing Dashboard

We deliver an interactive marketing dashboard for DrivePoints marketers to use to easily identify the best companies to target, and persist notes on companies in the dataset. This enables long-term use and the ability to fine-tune the model over time. Source code and supporting documentation for all code to run and make changes to the dashboard are available in the project's shared GitHub repository ([8]).

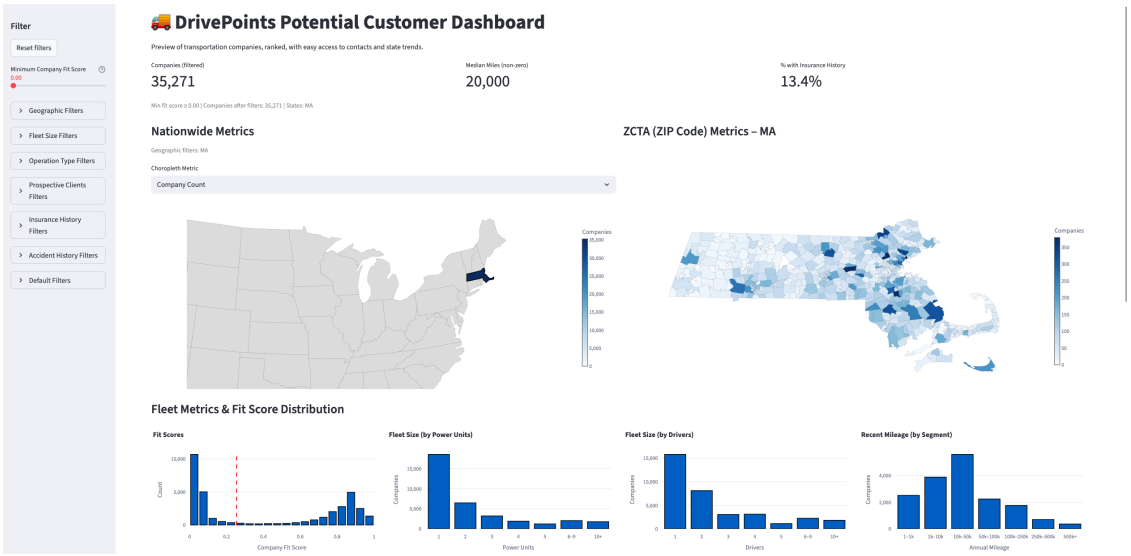

Figure 6: Screenshot of Streamlit Dashboard

4. Software Delivered

4.1 Jupyter Notebooks & Python Scripts

Source code and supporting documentation for all code are available in the project's shared GitHub repository ([8]).

We developed a modular data analysis pipeline in Python using Jupyter Notebooks to process, evaluate, and filter the FMCSA dataset. The codebase includes reusable scripts for data ingestion, preprocessing, data quality scoring, joining auxilliary datasets, llm experiments, and total addressable market (TAM) estimation. Each component is designed to operate independently, allowing DrivePoints or future researchers to update datasets or adjust scoring criteria without altering the full workflow.

All scripts are organized to ensure transparency and reproducibility. Intermediate data and outputs are checkpointed at each stage to support auditability and allow resumption from partial runs. The notebooks can be parameterized to run on updated monthly FMCSA data releases, enabling continuous refresh of the TAM and DQS metrics.

4.2 Interactive Streamlit Dashboard

Source code and supporting documentation for all code to run and make changes to the dashboard are available in the project's shared GitHub repository ([8]).

To make the tool directly actionable for DrivePoints' analysts, we implemented an interactive dashboard interface built with open-source Python frameworks. The primary front-end uses **Streamlit**, a lightweight and free platform that allows analysts to interact with the full dataset and logistic

regression outputs in real time without writing code. Streamlit components such as sliders, dropdowns, and multiselect filters provide a fast and intuitive way to adjust thresholds and explore subsets of the data.

The dashboard allows users to:

- **Adjust filtering thresholds** for geography, fleet size, DQS cutoff, cargo type, safety score, or logistic regression probability, observing changes to the TAM and Company Fit Scores interactively.
- **Keep track of DrivePoints' relationship to each company** by filling in the progress made with each company (e.g., noting that they are in talks with a company, are not interested, are a client, etc.).
- **Search particular companies and export company lists** using the continuous "Company Fit Score," filtering by state, fleet size, or safety performance.
- **Visualize results dynamically** through linked dashboards that include U.S. choropleth maps of TAM distribution, bar charts of model coefficients and ROC/F1 evaluation metrics for the logistic regression model.

5. Discussion

5.1 Conclusion

This project demonstrates that publicly available federal transportation data can be transformed into a reliable and scalable system for customer-targeted acquisition in the trucking insurance domain. Our analysis shows that the FMCSA Motor Carrier Census is sufficiently complete, stable over time, and structurally consistent to support downstream business decision-making when paired with explicit data quality controls.

While LLMs successfully identified certain subtle inconsistencies and implausible records, their classification behavior exhibited a strong bias toward negative labels and lower balanced accuracy. Combined with higher operational cost and reduced interpretability, these limitations make LLMs unsuitable as a primary mechanism for large-scale company fit assessment. Instead, they are best positioned as complementary tools for targeted audits or qualitative validation where contextual reasoning provides incremental value.

We implemented a statistical scoring framework based on logistic regression to estimate the probability that a company is a good fit for DrivePoints. The resulting Company Fit Score enables full-population ranking of carriers, allowing DrivePoints to identify high-quality prospects even when they do not satisfy every individual rule. Model performance, as measured by AUC, demonstrates strong ranking ability, and a structured feedback mechanism for iterative label refinement will lead to ongoing system improvement.

Taken together, this work delivers a production-ready targeting pipeline consisting of a validated data foundation, transparent eligibility rules, an interpretable probabilistic scoring model, and an operational dashboard. More broadly, it illustrates how public data can be transformed into actionable customer acquisition business intelligence.

5.2 Future Work

While this system was developed for DrivePoints, the underlying pipeline is intentionally generalizable. Any organization seeking customer-targeted acquisition from large administrative datasets can adopt this approach by defining its own notion of customer fit, annotating a modest sample of entities, and retraining the scoring model. Because data validation, eligibility filtering, and probabilistic ranking are modularized, each component can be adapted to different business objectives without re-engineering the full system.

Several natural extensions would further enhance the pipeline's accuracy and flexibility. First, expanding the labeled dataset would support richer feature interactions, improve model calibration, and enable more robust evaluation across subsegments of the trucking market. Second, large language models could be incorporated as a secondary validation layer to flag ambiguous, high-impact, or inconsistent records for targeted human review, strengthening the validity component of the Data Quality Score without introducing LLM bias into large-scale ranking.

Additional data enrichment opportunities also merit further exploration. Preliminary experiments with address geocoding using the U.S. Census geocoder ([12]) achieved partial coverage, suggesting that improved preprocessing or alternative geocoding services could enable spatial features for regional segmentation and market analysis. Similarly, integrating external business datasets such as the Infogroup Historical Business Database ([13]) yielded approximately 170,000 high-confidence matches with FMCSA records, providing NAICS classifications for a substantial subset of carriers. With further refinement, this linkage could support industry-level segmentation, NAICS prediction for unmatched firms, and more granular targeting strategies.

Together, these extensions would strengthen the pipeline's adaptability, support more nuanced market segmentation, and expand its applicability beyond the trucking insurance domain to other customer acquisition and underwriting contexts.

6. Team Contributions

- Ben Sullivan: Found auxiliary datasets to enrich the census data (insurance history data, FARS and CRSS data). Owned final Streamlit dashboard, leading development of all features, including the filtering systems, dynamic visualizations, and the read-write features. Categorized the 70,000+ unique tokens in the "cargo_carried" into 36 categories.
- Jared Donohue: Defined Data Quality Score (DQS) metric. Ran LLM experiments. Coordinated ground truth annotations for stat model training and validation. Fetched key

"cargo_carried" feature using API and joined to dataset. Led project management for team.

- Moacir P. de Sá Pereira: Investigated geocoding the dataset and then built the bridge to the Infogroup Historical Business Dataset. Collected and analyzed the historical census data. Organized and documented GitHub repository.

- Jialin Wen: Led the end-to-end development of the statistical modeling pipeline, integrating the rule-based Data Quality Score (DQS) computation into downstream modeling and market sizing workflows. Owned the TAM modeling and exploratory data analysis to inform feature selection and filtering criteria. Designed and implemented the logistic regression model, addressing high-cardinality categorical features through target encoding and optimizing the modeling pipeline to improve scalability and efficiency. Conducted comprehensive model evaluation and threshold analysis, ultimately establishing a closed-loop framework that connects data quality assessment with high-quality company fit prediction.

References

[1] DrivePoints. *DrivePoints Insurance Technology Overview*. Company overview describing fleet management and insurance optimization platform. Accessed October 2025. 2025. URL: https://drivepoints.com.

[2] American Trucking Associations. *Economics & Industry Data*. https://www.trucking.org/economics-and-industry-data. Accessed October 30, 2025. 2025.

[3] Federal Motor Carrier Safety Administration. *Comprehensive Safety Analysis (CSA) Monthly Data Run*. Includes registration data for active interstate and intrastate hazmat motor carriers of property and passengers. September 15, 2025 update. 2025. URL: https://ai.fmcsa.dot.gov/sms/.

[4] Federal Motor Carrier Safety Administration. *US DOT Insurance History Data*. Dataset of previous insurance policies for carriers, brokers, and freight forwarders. Updated September 24, 2025. 2025. URL: https://ai.fmcsa.dot.gov/insurance/.

[5] National Highway Traffic Safety Administration. *Fatality Analysis Reporting System (FARS)*. National database listing every fatal crash in the U.S. Accessed October 2025. 2025. URL: https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars.

[6] National Highway Traffic Safety Administration. *Crash Report Sampling System (CRSS)*. National sample of U.S. police-reported crashes. Accessed October 2025. 2025. URL: https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system.

[7] Federal Motor Carrier Safety Administration. *QCMobile API Developer Documentation*. Developer documentation for the QCMobile API, which provides programmatic access to U.S. DOT motor carrier registration, insurance, authority, and safety performance data. Accessed December 17, 2025. 2025. URL: https://mobile.fmcsa.dot.gov/QCDevsite/docs/qcApi.

[8] *DSI Capstone 2025: Trucking Market Research (DrivePoints Insurance)*. GitHub repository. 2025. URL: https://github.com/drivepoints/dsi-2025-fall-team-35-trucking-market-research.

[9] Federal Motor Carrier Safety Administration. *Safety Measurement System (SMS) Motor Carrier Census Data*. Version 108. Version 108, retrieved October 15, 2025. Contains registration and safety data for active U.S. motor carriers. Oct. 2025. URL: https://ai.fmcsa.dot.gov/SMS/.

[10] scikit-learn developers. *sklearn.metrics.balanced_accuracy_score*. scikit-learn. 2024. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html (visited on 11/02/2025).

[11] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[12] United States Census Bureau. *Census Geocoder*. URL: https://geocoding.geo.census.gov/geocoder/ (visited on 10/08/2025).

[13] Data Axle. *InfoGroup Historical Business Data*. Annual dataset providing geolocated U.S. business listings with NAICS codes. 2025. URL: https://www.data-axle.com/solutions/data-axle-business-data/.