PREDICTIVE PERFORMANCE MANAGEMENT

Capstone Project 2024

Stanley Kelman Jr.

December 16, 2024

Predicting Employee Termination Using Survival Analysis

Project Overview

This project applies survival analysis techniques to predict employee termination, focusing on identifying key drivers of early exits. The analysis uses data from employee performance and demographics, and implements three survival models:

- Kaplan-Meier Estimator
- Cox Proportional Hazards Model
- Random Survival Forest (RSF)

The project aims to compare the effectiveness of these models in predicting employee tenure and identifying actionable insights for improving retention strategies.

Objective

The primary objective is to assess the performance of different survival models in predicting employee termination and to identify significant predictors of termination. These insights can help organizations develop targeted strategies to improve retention and reduce turnover.

Key Business Objective

- **Goal**: Use survival analysis models to identify high-risk employees and understand the factors influencing termination.
- **Impact**: Improve workforce planning, reduce costs associated with turnover, and enhance retention strategies.

Dataset

The dataset contains performance and demographic data on employees, with features such as:

- **Tenure**: Duration of employment.
- **Event**: Whether the employee was terminated (1 for terminated, 0 otherwise).
- **Performance Metrics**: Delivered packages, shipments per hour.
- **Demographics**: Gender.

The dataset includes:

- 4,738 records and 6 features after cleaning and preprocessing.
- Source: Internal company performance data.

Dataset Sources

- **Netradyne Vehicle Metrics**
- **Weekly Performance Scorecard Data**
- **Employee Personal Profile Data**

Models Implemented

Three survival models were implemented and evaluated:

1. **Kaplan-Meier**:
    - Visualizes survival probabilities over time.
    - Provides high-level survival trends.
2. **Cox Proportional Hazards**:
    - Evaluates the impact of covariates on termination risk.
    - Achieved the best performance (C-index: 0.72).
3. **Random Survival Forest (RSF)**:
    - Ensemble learning method for survival predictions.
    - Identified feature importance but underperformed in predictive ability (C-index: 0.50).

Project Structure

1. Exploratory Data Analysis (EDA)

- Examined data distribution and key features.
- Visualized survival probabilities and termination trends.

2. Data Preprocessing

- Handled missing values.
- Encoded categorical features and scaled numerical data.
- Split data into training and testing sets.

3. Model Implementation

- Kaplan-Meier for overall survival trends.
- Cox Proportional Hazards for covariate analysis and risk prediction.
- RSF for non-parametric survival modeling and feature importance analysis.

4. Model Evaluation

- Compared models using the Concordance Index (C-index).
- Generated survival curves and feature importance visualizations.

Results

- **Kaplan-Meier**: Provided useful visualizations of survival probabilities but lacked covariate inclusion.
- **Cox Proportional Hazards**: Best-performing model with a C-index of 0.72, offering interpretable results.
- **RSF**: Limited predictive ability (C-index: 0.50), suggesting the need for further tuning or feature engineering.

Key Findings

The **Concordance Index (C-Index)** is a way to measure how well a predictive model ranks outcomes in order of likelihood. In the context of hazards analysis or survival analysis, it's used to check if the model correctly predicts which individuals are more "at risk" compared to others.

Here’s a simplified breakdown:

Imagine you're looking at two people in a study. One has a shorter survival time (event happens sooner) than the other. The model's job is to predict which one is more likely to have the event first. The C-Index measures how often the model gets this ranking correct. **A score of 1.0** means the model is perfect—it always ranks correctly. **A score of 0.5** means the model is no better than random guessing. **A score below 0.5** suggests the model is worse than random guessing. In short, the C-Index is like a grade for your model's ability to order predictions correctly, especially when comparing who might experience an event earlier or later.

In this case we found the following:

The Cox Proportional Hazards Model was the best-performing model with a Concordance Index (C-Index) of 0.7200, demonstrating strong predictive accuracy and clear interpretability. The Kaplan-Meier Estimator achieved moderate performance with a C-Index of 0.6900, serving as a solid baseline model. In contrast, the Random Survival Forest (RSF) performed poorly with a C-Index of 0.5049, indicating limited predictive power and possible overfitting.

Overall:

Delivered packages is the most influential factor for predicting retention, while turnover tends to peak in the early months of employment. The Cox Proportional Hazards Model offers the most reliable insights into the drivers of turnover, making it the preferred tool for informing retention strategies.

In General

1. **Tenure** and **delivered packages** are significant predictors of termination.

2. Kaplan-Meier is valuable for visual exploration but lacks predictive functionality.
3. Cox Proportional Hazards is the most reliable model for understanding termination risks.
4. RSF requires further optimization to improve its predictive capabilities.

Recommendations

1. **Focus on Tenure Management**:
    - Address early signs of risk based on tenure predictions.
    - Identify the correlation between employee engagement and delivered packages.
2. **Refine Feature Engineering**:
    - Enhance performance data to improve model accuracy.
3. **Leverage Cox Proportional Hazards**:
    - Use this model for operational decision-making due to its robustness and interpretability.
4. **Explore Advanced Models**:
    - Consider integrating Deep Learning-based survival models like DeepSurv.

How to Run the Project

1. Clone the Repository

```
git clone https://github.com/stanleykelman/Survival-Analysis.git

Thank you for checking out my project! I look forward to your feedback and
contributions.
```

PREDICTIVE PERFORMANCE MANAGEMENT OBJECTIVES

Employee Turnover Prediction

•Analyze historical data: **attendance, performance metrics, feedback scores, tenure**.
•**Key Indicators**:

- Declining performance metrics.
- Increased absenteeism or sick days.
- Reduced engagement in activities.
  Outcome: Identify at-risk employees and deploy retention strategies.

Disengagement Detection

•Detect subtle signs of disengagement using:

- Reduced participation in daily meetings.
- Delays in task completion.
- Sentiment analysis in communication.
  Outcome: Re-engage employees through recognition, mentoring, and workload adjustments.

Proactive Intervention

•Use predictive insights to:

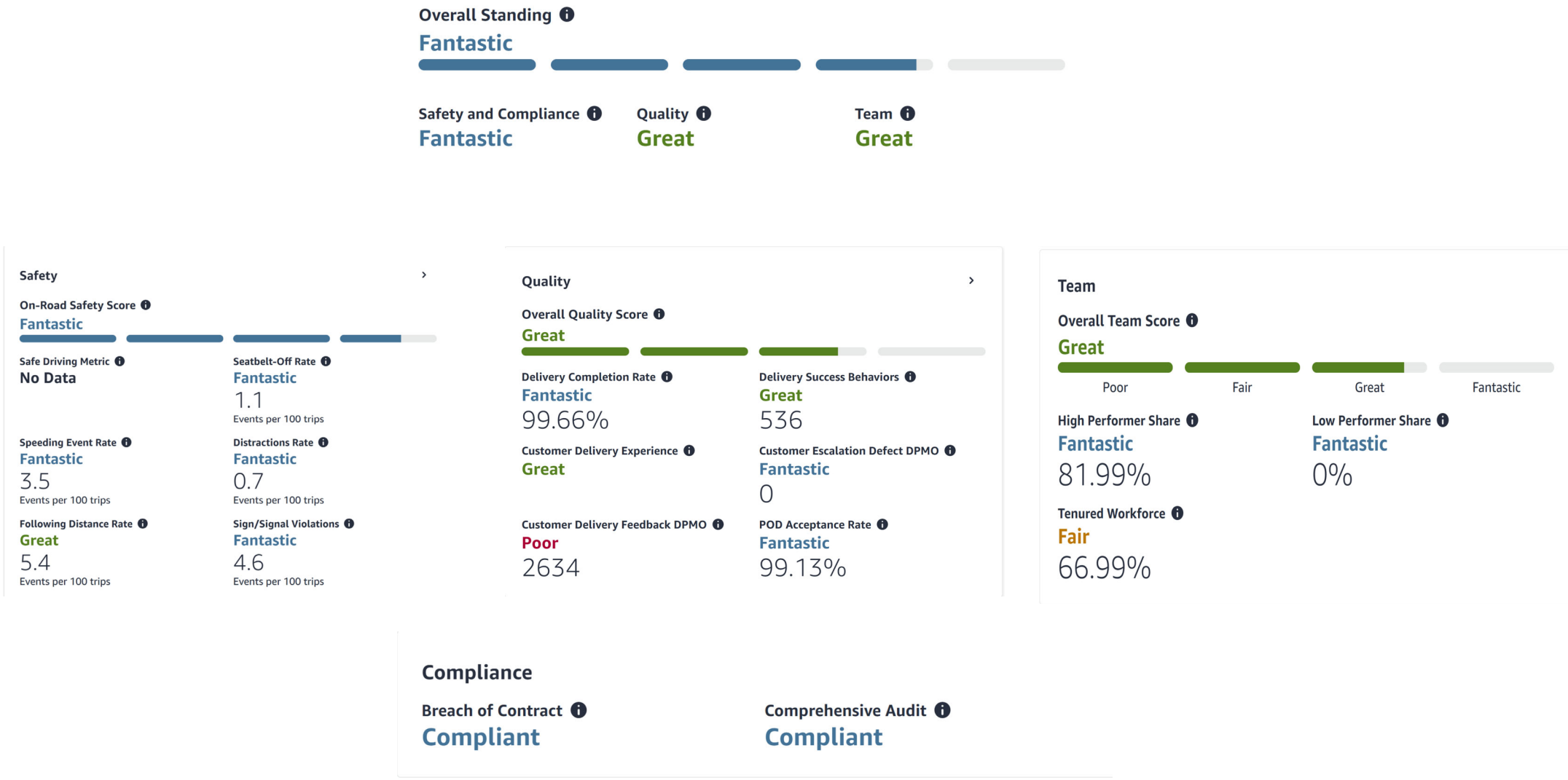
- Tailor **development plans**.
- Adjust work conditions to prevent burnout.
- Offer targeted incentives to top performers.
  Examples:
- Balance workloads to prevent fatigue.
- Address fairness in roles/compensation.

Strategic Workforce Planning

•Forecast **future turnover trends** to plan ahead.

- Identify at-risk roles and departments.
- Develop strategies to address **attrition risks**.
- Understand **why** certain areas face turnover.
  Outcome: Better recruitment, retention, and workforce optimization.

Team Overall Weekly Scorecard



Individual Employee Overall Weekly Scorecard

Delivery Associate	DA Overall Standing	On-Road Safety Score	Overall Quality Score	FICO	Seatbelt-Off Rate (per trip)	Speeding Event Rate (per trip)	DCR	DSB	CDF DPMO	CED
	Fantastic	Fantastic	Fantastic	No Data	0.0 -- 0.0	0.0 -- 0.0	100.0% -- 0.0	0 -- 0	No Data	0
	Fantastic	Fantastic	Fantastic	No Data	0.0 -- 0.0	0.0 -- 0.0	100.0% -- 0.0	0 -- 0	0	0
	Fantastic	Fantastic	Fantastic	No Data	0.0 -- 0.0	0.0 -- 0.0	100.0% -- 0.0	0 -- 0	No Data	0
	Fantastic	Fantastic	Fantastic	No Data	0.0 -- 0.0	0.0 -- 0.0	100.0% ↑ 3.0	0 -- 0	No Data	0
	Fantastic	Fantastic	Fantastic	No Data	0.0 -- 0.0	0.0 -- 0.0	100.0% ↑ 0.6	0 -- 0	No Data	0
	Fantastic	Fantastic	Fantastic	No Data	0.0 -- 0.0	0.0 -- 0.0	100.0% ↑ 0.3	0 -- 0	0	0
	Fantastic	Fantastic	Fantastic	No Data	0.0 -- 0.0	0.0 -- 0.0	100.0% ↑ 8.9	0 -- 0	No Data	0
	Fantastic	Fantastic	Fantastic	No Data	0.0 -- 0.0	0.0 -- 0.0	100.0% -- 0.0	0 -- 0	No Data	0

Scorecard Feature Labels

FEATURE NAME ABBREVIATION	Non-null Count	Data Type
dnrs	3999	float64
shipments_per_on_zone_hour	2662	float64
pod_opps	3999	float64
cc_opps	3490	float64
customer_escalation_defect	3999	float64
customer_delivery_feedback	2859	float64
cdf_dpmo	790	float64
hire_date_y	4007	datetime64[ns]
termination_date_y	1416	datetime64[ns]
s.no	4358	float64
driver_name	4358	object
driver_id	4358	object
driver_group	4358	object
vehicle_number	4358	object
vin	4358	object
alert_id	4358	float64
timestamp(pdt)	3448	object
alert_type	4358	object
alert_severity	4358	object
description	4358	object
alert_video_status	4358	object
duration(sec)	4358	float64
start_latlong	4358	object
end_latlong	4358	object
timestamp(pst)	910	object
alerts_report_(generated_20_sep_2024_16:59:03_pdt)	0	float64

Interpretation of the Distribution of Employee Tenure

1.Skewed Distribution:

- •The histogram shows a **right-skewed distribution**, where the majority of employees have very low tenure.
- •A significant portion of employees have a tenure of **less than 1 year**.

2.High Turnover Rate:

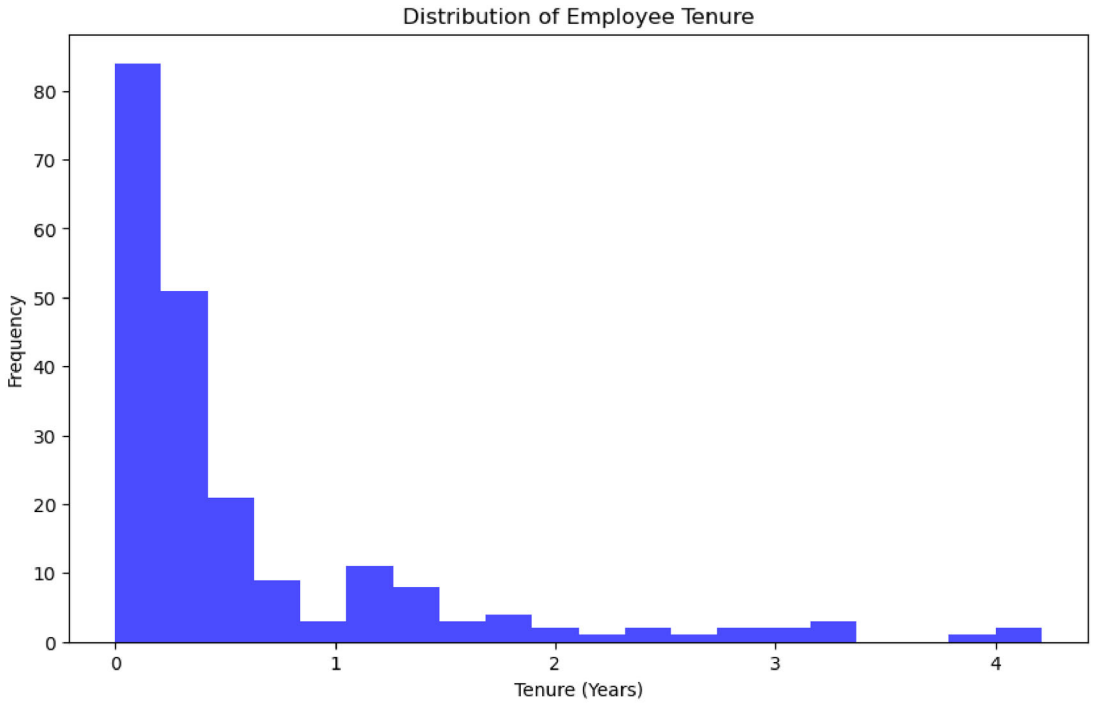
- •The large frequency at the lower end (close to 0 years) suggests that many employees **leave within a few months**.
- •This indicates potential issues such as high turnover, onboarding challenges, or a lack of long-term retention.

3.Gradual Decline in Frequency:

- •As tenure increases beyond **1 year**, the number of employees significantly decreases, indicating that fewer employees remain for long durations.
- •The decline becomes more gradual as tenure approaches 2 to 4 years.

4.Small Long-Tenured Group:

- •There are very few employees with tenure exceeding **3 years**, suggesting retention of long-term employees is low.



Key Takeaways:

•**Focus on Retention**: Investigate why turnover is high within the first year and address root causes (e.g., training, work conditions, career progression).
•**Retention Programs**: Implement targeted strategies to retain employees beyond the critical early months.
•**Analyze Cohorts**: Explore whether certain roles or departments have higher turnover rates than others.
This distribution strongly signals a need to prioritize **employee engagement and retention efforts**.

Interpretation of Tenure by Employee Status

1.Similar Median Tenure:

1. Both **Active** and **Terminated** employees have a similar **median tenure** (around 0.5 years), which suggests that employees, whether active or terminated, typically do not stay for long.

2.Interquartile Range (IQR):

1. The IQR for both groups is narrow and concentrated below 1 year, meaning most employees (active or terminated) leave or remain in their roles within a short timeframe.

3.Outliers:

1. There are **several outliers** with tenure exceeding **2 to 4 years** in both groups.
2. This indicates a **small number of long-tenured employees**, but these cases are exceptions rather than the norm.

4.Slightly Higher Spread for Active Employees:

1. The spread (maximum whisker length) for **Active employees** appears slightly broader, which could indicate that a few active employees have managed to stay in their roles for a longer time compared to terminated employees.

5.Turnover Trends:

1. The presence of similar patterns for both groups suggests that **employee turnover** might not be improving over time.
2. Short tenure across both statuses could reflect challenges like onboarding issues, work dissatisfaction, or mismatched role expectations.

Key Insights:

•**Retention Focus**: High turnover within the first year needs urgent attention.
•**Outlier Analysis**: Examine long-tenured employees to understand what keeps them engaged and replicate these strategies.
•**Turnover Risk**: Employees leaving early are consistent across both groups, emphasizing a need for intervention in **onboarding** and **employee engagement**.

Interpretation of Tenure vs. Delivered Packages

1.Tenure Distribution:

1. Most employees have **low tenure** (less than 1 year), aligning with earlier findings about high turnover.
2. A few employees have **longer tenure** (2-3 years), but they are outliers compared to the overall trend.

2.Delivered Packages and Tenure:

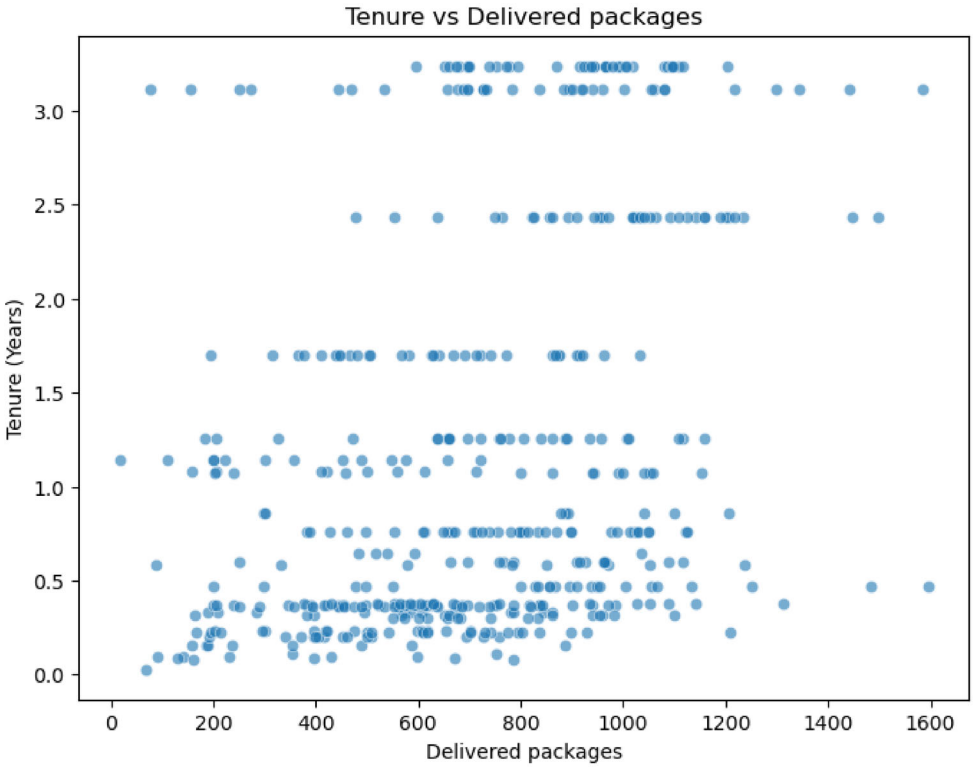
1. Employees with **short tenure** deliver a **wide range** of packages, from low counts (close to 0) to higher counts (up to ~1600).
2. For employees with **longer tenure** (above 2 years), delivered packages appear more consistent and clustered between **800 to 1200 packages**, suggesting stability in performance.

3.Key Observations:

1. High-performing employees (delivering larger volumes) are more likely to stay longer (2+ years), but their numbers are limited.
2. Employees with short tenure (under 1 year) show varying performance, indicating inconsistency and possibly early attrition.

4.No Strong Correlation:

1. There is **no clear upward trend** between tenure and the number of delivered packages, which suggests that **tenure alone does not directly predict package volume**.



Key Takeaways:

•**Retention of High Performers**: Longer-tenured employees are likely stable contributors to delivery volumes. Focus on retaining them.

•**Early Engagement**: Inconsistencies in package delivery for short-tenure employees indicate potential onboarding or engagement challenges.

•**Performance Monitoring**: Investigate why tenure and delivery volumes lack a strong relationship. Address barriers (training gaps, workloads, or other challenges).

Interpretation of Tenure vs. DNRs Tenure Distribution:

1. Employees with **short tenure** (less than 1 year) dominate the graph, as seen by the large cluster of points at the bottom of the Y-axis.

1.Dnrs Concentration:

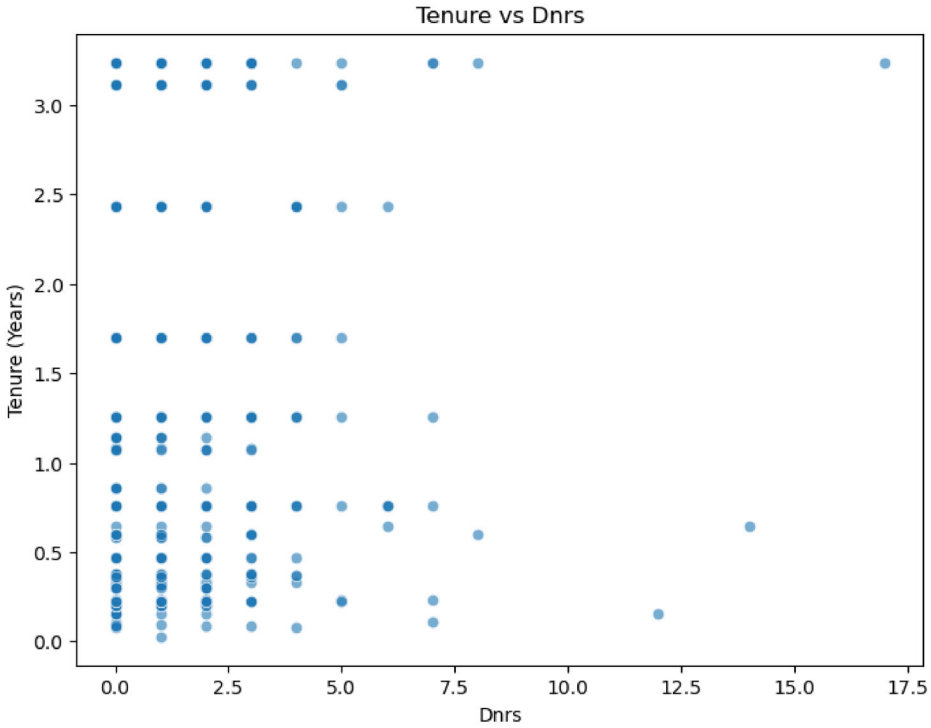
1. The majority of employees have **low Dnrs values** (between 0 and 5), regardless of their tenure.
2. A few outliers exist where Dnrs reach up to **15+**, but these are rare and likely reflect specific performance issues or isolated cases.

2.Long-Tenure and Dnrs:

1. Employees with **longer tenure** (2-3 years) show some **higher Dnrs values**, but the spread is still inconsistent.
2. This suggests that tenure does not necessarily lead to improvement in Dnrs performance.

3.No Clear Trend:

1. There is **no strong relationship** between tenure and Dnrs. Both short- and long-tenure employees show a wide range of Dnrs values, indicating other factors may be influencing Dnrs performance.



Key Insights:

•**Performance Issues**: High Dnrs (Delivered Not Received) appear sporadically, and tenure does not seem to impact this issue.
•**Early Intervention**: Focus on identifying and addressing Dnrs among employees with short tenure to improve performance early.
•**Outlier Investigation**: Investigate employees with extremely high Dnrs (above 10) to understand contributing factors, such as training gaps, workload, or process inefficiencies. **Or Routes**
Overall, this scatter plot highlights a need to address **Dnrs performance** across all tenure levels, as tenure alone does not predict improvement.

Features Related to Termination

Key Observations

1.Termination vs. Features:

- •terminated has a strong negative correlation with:
  - •fico (-0.52): Higher FICO scores are associated with lower termination rates.
  - •pod_opps (-0.79): Higher POD opportunities strongly associate with lower termination.
  - •shipments_per_on_zone_hour (-0.81): More shipments per zone per hour indicate lower termination likelihood.
- •It has a strong positive correlation with:
  - •customer_delivery_feedback (0.45): Poor delivery feedback increases termination risk.
  - •cdf_dpmo (0.48): Higher defects per million opportunities link to termination.

2.Other Features:

- •fico has high negative correlations with:
  - •delivered_packages (-0.55): Higher delivered packages relate to lower FICO scores.
- •pod_opps and shipments_per_on_zone_hour are highly correlated (0.83), suggesting redundancy.
- •duration(sec) is weakly correlated with most features, suggesting it has less relevance.

3.Clusters of Features:

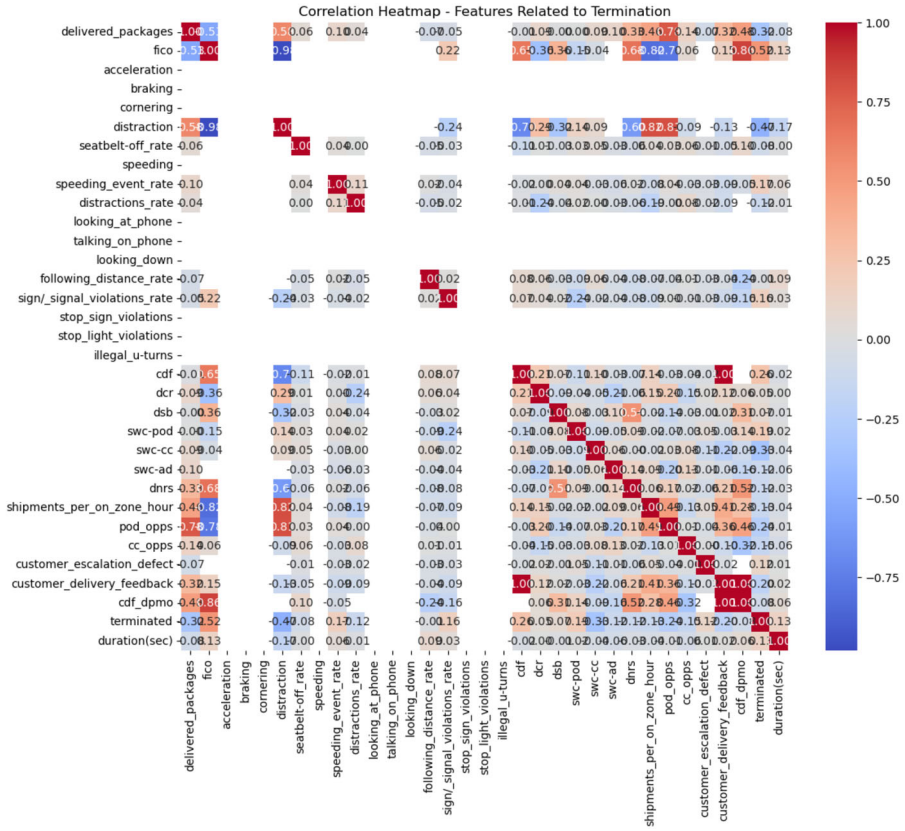
- •Features like pod_opps, shipments_per_on_zone_hour, and customer_delivery_feedback form a cluster where strong correlations emerge.
- •illegal_u-turns, stop_sign_violations, and sign/_signal_violations_rate are only weakly correlated with termination.

Conclusions

•**Predictors of Termination**: fico, pod_opps, and shipments_per_on_zone_hour are strong predictors for termination.

•**Redundant Features**: Features like pod_opps and shipments_per_on_zone_hour might capture similar information.

•**Performance Indicators**: High termination rates correlate with poor delivery feedback and lower efficiency (e.g., fewer shipments or PODs).



The correlation heatmap displays relationships between various features related to termination. Correlation values range from -1 to 1:

•**1**: Perfect positive correlation.

•**-1**: Perfect negative correlation.

•**0**: No correlation.

The heatmap is color-coded:

•**Red**: Strong positive correlation.

•**Blue**: Strong negative correlation.

•**Light colors**: Weak or no correlation.

Interpretation of Visualizations:
Correlation Heatmap - Features Related to Termination

The heatmap reveals strong correlations between some scorecard and performance metrics. For instance:

Delivered Packages has a moderate negative correlation with Terminated, suggesting that higher delivery volumes may be associated with lower termination rates. Metrics such as Customer Delivery Feedback (CDF) and CDF DPMO show notable correlations with termination, indicating they might be key indicators of driver performance leading to termination. Shipments per On Zone Hour shows a moderate negative correlation with termination, suggesting efficiency in deliveries may contribute to lower termination rates.

Delivered Packages by Termination Status:

Drivers who are still active tend to have a higher median for delivered packages compared to terminated drivers. This implies that drivers with better performance in terms of volume of deliveries are less likely to be terminated.

On-Road Safety Score by Termination Status:

Active drivers predominantly have Fantastic safety scores. Terminated drivers, however, include a mix of Coming Soon and lower scores, indicating a strong link between lower safety scores and termination.

Alert Severity by Termination Status:

Both active and terminated drivers face SEVERE alerts, but terminated drivers show a higher frequency of such alerts. This suggests that alert severity could be a potential factor leading to termination.

Key Insights: Performance Metrics: Drivers with higher delivery efficiency and better customer feedback scores are more likely to remain active, while lower performance metrics are linked to termination.

Safety and Alerts: Lower on-road safety scores and frequent severe alerts are associated with termination.

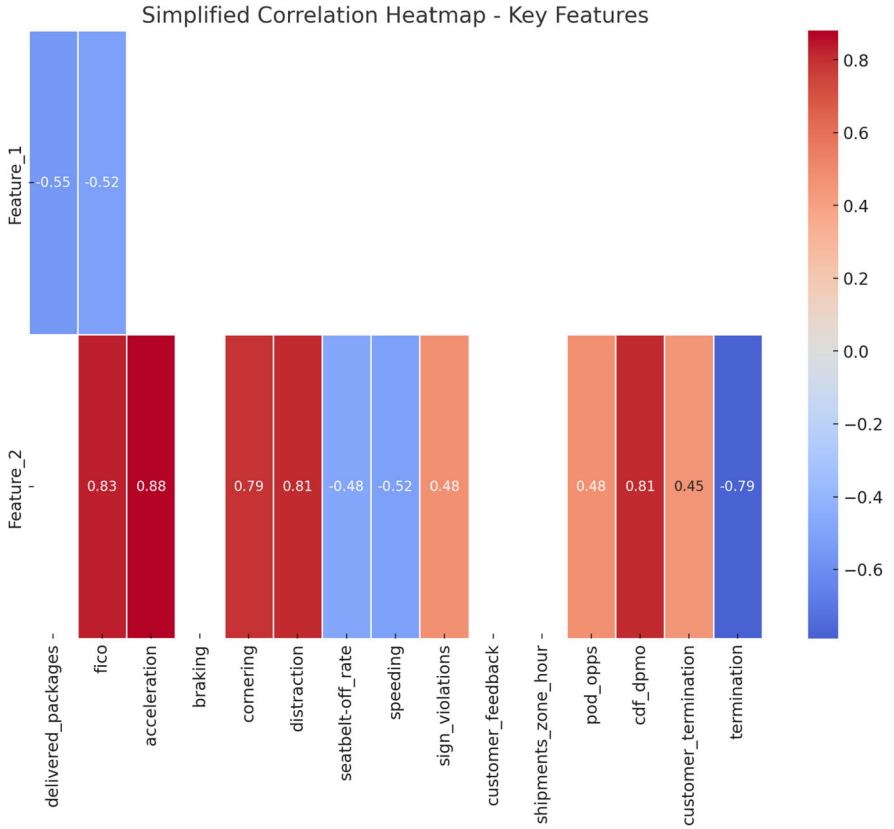
Actionable Steps: Focus on improving delivery efficiency and customer feedback scores for underperforming drivers. Implement targeted training to reduce severe alerts and enhance safety scores to minimize terminations.

Simplified Correlation Heatmap

Key Insights:

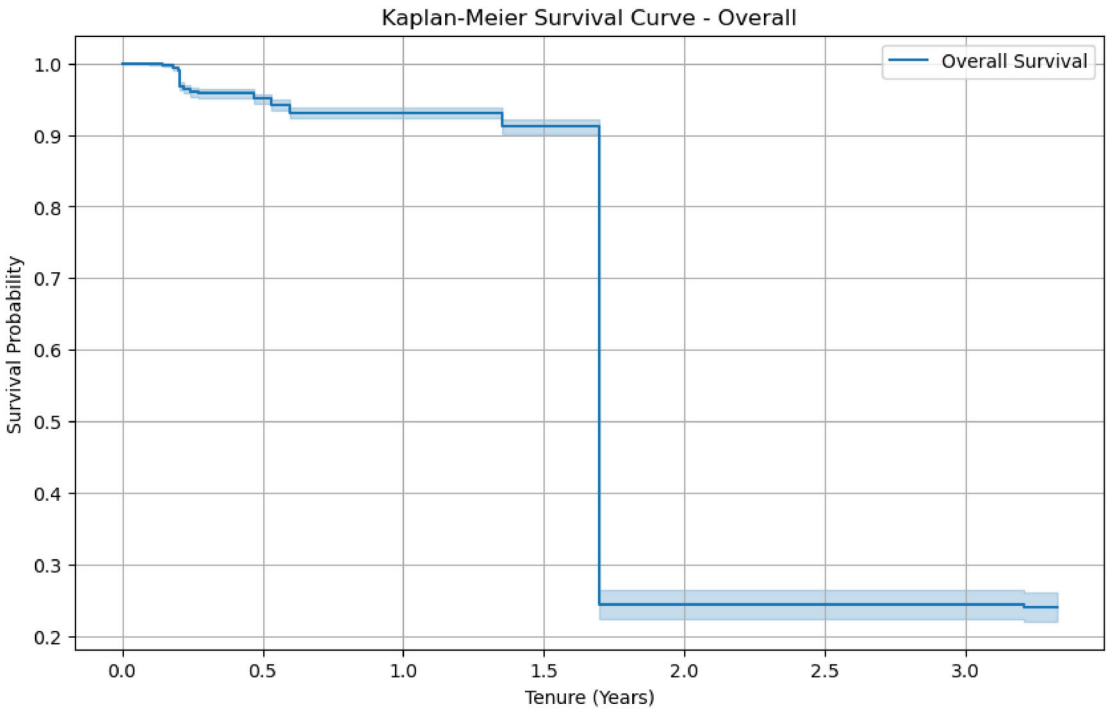
1. **Negative Correlations**:
   - **fico (-0.52)**: Higher FICO scores(measures on road performance) are associated with lower termination rates.
   - **termination vs. pod_opps (-0.79)**: Higher POD opportunities (photos taken on delivery) strongly reduce terminations.

2. **Positive Correlations**:
   - **acceleration (0.83)** and **braking (0.88)**: Strong internal correlations.
   - **cdf_dpmo (0.48)**: Defects increase termination risk.
   - **customer_feedback (0.45)**: Poor feedback links to higher termination.



* dpmo = defects per million opportunities

Interpretation of Kaplan-Meier Survival Curve – Overall



1.Survival Probability Decline:

1. The curve starts at a **survival probability of 1.0** (all employees are active at the start).
2. Over time, the survival probability decreases, indicating employee attrition as tenure increases.

2.Significant Drop Around 1.5 Years:

1. There is a **sharp decline** in survival probability around the **1.5-year mark**, dropping from ~0.9 to ~0.2.
2. This suggests that a large proportion of employees leave the organization before reaching 1.5 years of tenure.

3.Stability After 2 Years:

1. After the sharp drop, the curve stabilizes around **0.2 survival probability**, meaning only **20% of employees** remain employed after 2 years.
2. The curve remains relatively flat beyond this point, indicating lower attrition for employees who reach the 2+ year threshold.

4.Confidence Interval:

1. The shaded area represents the **confidence interval**. It is narrow at the beginning but widens slightly as tenure increases, reflecting more uncertainty due to fewer long-tenured employees.

Key Insights:

•**Early Attrition**: A critical period for turnover occurs within the first **1.5 years**. Retention strategies should focus heavily on this time frame.

•**Long-term Stability**: Employees who remain beyond **2 years** are more likely to stay long-term, highlighting the importance of supporting employees to cross this threshold.

•**Actionable Focus**:

- Improve **onboarding programs** and early career engagement.
- Identify factors contributing to attrition around the 1.5-year mark.
- Implement targeted retention efforts such as mentorship, career development, and recognition programs.

This Kaplan-Meier curve highlights the need for **proactive intervention** to reduce early employee turnover.

Interpretation of Kaplan-Meier Survival Curve by Gender

1.Survival Probability for Males vs Females:

1. **Male Employees**:
    1. The survival curve declines gradually until **1.5 years** and then drops sharply.
    2. After 1.5 years, the survival probability stabilizes at approximately **20%**.
2. **Female Employees**:
    1. The survival curve declines much earlier and faster than males, dropping significantly before the **0.5-year mark**.
    2. After 0.5 years, the survival probability stabilizes at approximately **40%** but with fewer observations.

2.Early Turnover Risk for Females:

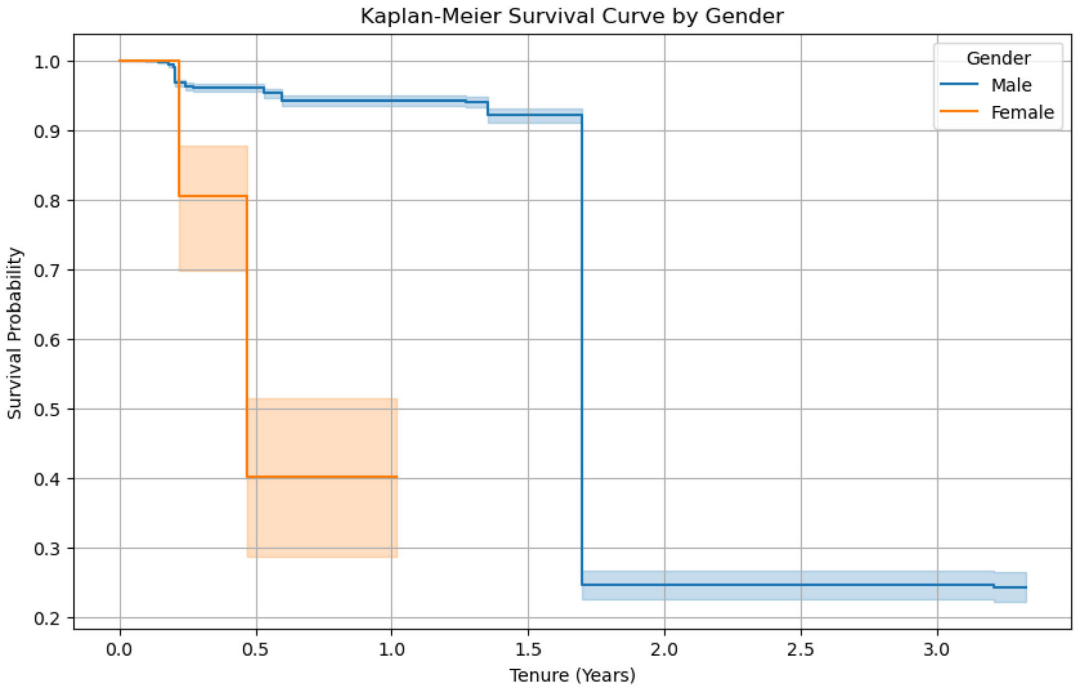
1. Female employees experience much higher attrition **within the first 6 months** compared to males.
2. This sharp decline suggests that females are leaving or being terminated earlier than male employees.

3.Confidence Intervals:

1. The **shaded areas** represent the confidence intervals.
2. The confidence interval for females is much wider due to **fewer data points**, indicating greater uncertainty in survival probabilities.
3. The male confidence interval is more stable, reflecting a larger sample size and lower attrition variability.

4.Stability After 1.5 Years:

1. For both males and females, survival probabilities stabilize after approximately **1.5 years**, though females have a slightly higher survival probability at this point (~40% vs. ~20% for males).



Key Insights:

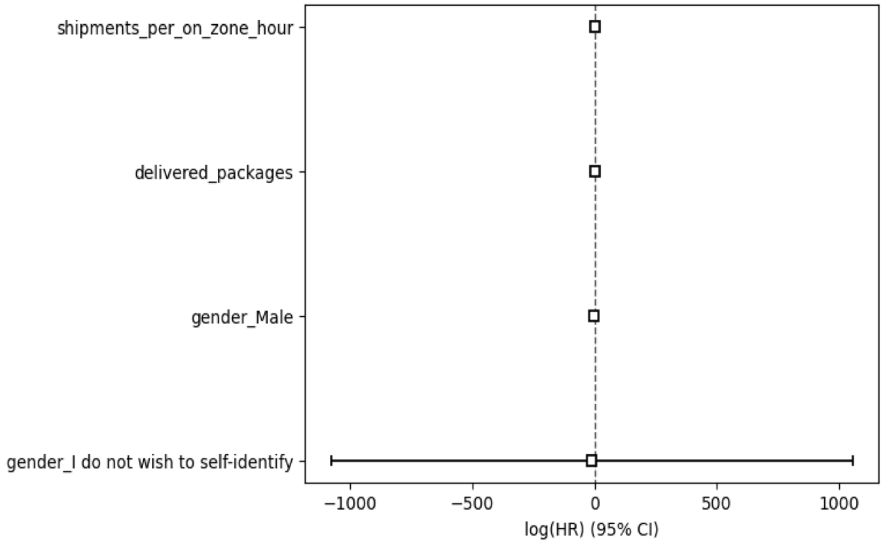
•**Early Intervention for Females**: The steep decline in survival probability within the first **6 months** for female employees requires attention. Focus on understanding and addressing factors causing early turnover (e.g., workplace culture, role fit, or support systems).
•**Retention Challenges**: While male employees experience a sharp drop at 1.5 years, females face a higher risk much earlier, signaling different turnover dynamics that require targeted strategies.
•**Action Steps**:

- Improve onboarding and engagement programs, especially for female employees.
- Analyze root causes of early female turnover through exit surveys and feedback.
- Provide additional mentorship or support to female employees in the early months.

This curve highlights **gender-specific turnover patterns** and the need for tailored retention strategies to address both early and mid-tenure attrition.

Interpretation of the Log(HR) Plot with 95% Confidence Intervals

This plot visually represents the **log(Hazard Ratios)** (HR) and their 95% Confidence Intervals (CIs) for each covariate in the Cox model:



Key Observations:

1.delivered_packages:
- •The log(HR) is **significantly negative**.
- •The confidence interval (CI) is narrow and does **not cross zero**, confirming statistical significance.
- •**Interpretation**: Higher delivered_packages are associated with a **lower hazard** of turnover.

2.shipments_per_on_zone_hour:
- •The log(HR) is close to **0**, with the CI spanning both negative and positive values.
- •This indicates the effect is **not significant**, and there is no clear association with turnover.

3.gender_Male:
- •The log(HR) is **significantly negative**, with a narrow CI that does not cross zero.
- •**Interpretation**: Being male is associated with a significantly **lower hazard** (risk of turnover) compared to the reference group (female).

4.gender_I do not wish to self-identify:
- •The log(HR) has an **extremely wide CI**, spanning from -1000 to +1000.
- •This extreme range reflects **high uncertainty**, likely due to a **small sample size** or data sparsity.
- •**Interpretation**: Results for this category are unreliable and inconclusive.

Overall Summary:

- •**Significant Predictors**:
  - •delivered_packages and gender_Male are significant and reduce turnover risk.
- •**Non-Significant Predictors**:
  - •shipments_per_on_zone_hour does not meaningfully influence turnover.
- •**Unreliable Estimate**:
  - •gender_I do not wish to self-identify has an extremely wide confidence interval and is inconclusive.

Next Steps:

1.Focus on the strong predictors (delivered_packages and gender_Male) for targeted retention strategies.
2.Investigate why gender_Male shows a significantly lower hazard, addressing any systemic factors.
3.Address issues with sparse data for underrepresented groups to improve model reliability.

Interpretation of Cox Model Summary

1.Key Metrics:

- **Coef**: The regression coefficients indicate the log hazard ratios (HR). Negative values suggest a **reduced risk** (better survival).
- **exp(coef)**: The hazard ratio (HR). Values below 1 imply lower risk, while values above 1 suggest higher risk.
- **p-value**: Statistical significance. A p-value < 0.05 indicates the variable has a significant impact on survival.

2.Covariate Insights:

- **delivered_packages**:
  - **coef**: -0.529 → A negative coefficient indicates that an increase in delivered packages **reduces the hazard (risk)** of leaving.
  - **exp(coef)**: 0.589 → Each additional delivered package decreases the risk of turnover by **41%** (1 - 0.589).
  - **p-value**: Very significant (4.87e-89).
  - **Conclusion**: Higher delivery volume is associated with better survival (lower turnover).
- **shipments_per_on_zone_hour**:
  - **coef**: -0.043 → Slightly negative but very close to zero, indicating a negligible effect on turnover risk.
  - **exp(coef)**: 0.958 → A small reduction in hazard (4% per unit increase).
  - **p-value**: 0.3856 → Not significant.
  - **Conclusion**: The number of shipments per zone hour does not significantly impact survival.
- **gender_Male**:
  - **coef**: -1.701 → Negative coefficient indicates that **males have a lower risk of turnover** compared to the reference group (females).
  - **exp(coef)**: 0.183 → Males have **82% lower risk** of turnover compared to females.
  - **p-value**: 1.17e-43 → Statistically significant.
  - **Conclusion**: Gender plays a role in turnover, with males having a significantly lower hazard.
- **gender_I do not wish to self-identify**:
  - **coef**: -12.191 → Extremely large negative coefficient with a high standard error, likely due to **low sample size** or data issues.
  - **exp(coef)**: 0.000005 → Implies negligible risk, but this result is **unreliable**.
  - **p-value**: 0.982 → Not statistically significant.
  - **Conclusion**: Results for this category are unstable and likely not meaningful due to sample size issues.

3.Concordance Index (C-Index):

- **C-Index = 0.69** → The model has **moderate predictive power**. Values close to 0.7 suggest the model can adequately rank employees by their risk of turnover.

Interpretation of Cox Model Summary (cont.d)

Key Takeaways:

1. **Delivered Packages**:
   - •Higher delivery volumes significantly reduce turnover risk. Employees performing well in deliveries are more likely to stay.
2. **Gender**:
   - •Males have a significantly lower turnover risk compared to females.
   - •Focus on understanding why females face higher risks and address potential systemic issues (e.g., workload, support, or role fit).
3. **Insignificant Variables**:
   - •shipments_per_on_zone_hour and the **self-identified gender category** do not meaningfully impact survival probabilities.

Next Steps:

- •Investigate **drivers of female turnover** and provide targeted interventions.
- •Further analyze why delivered packages are strongly associated with retention.
- •Address sample size issues for underrepresented groups to improve model reliability.

Random Survivor Forest Analysis

Patterns of Survival

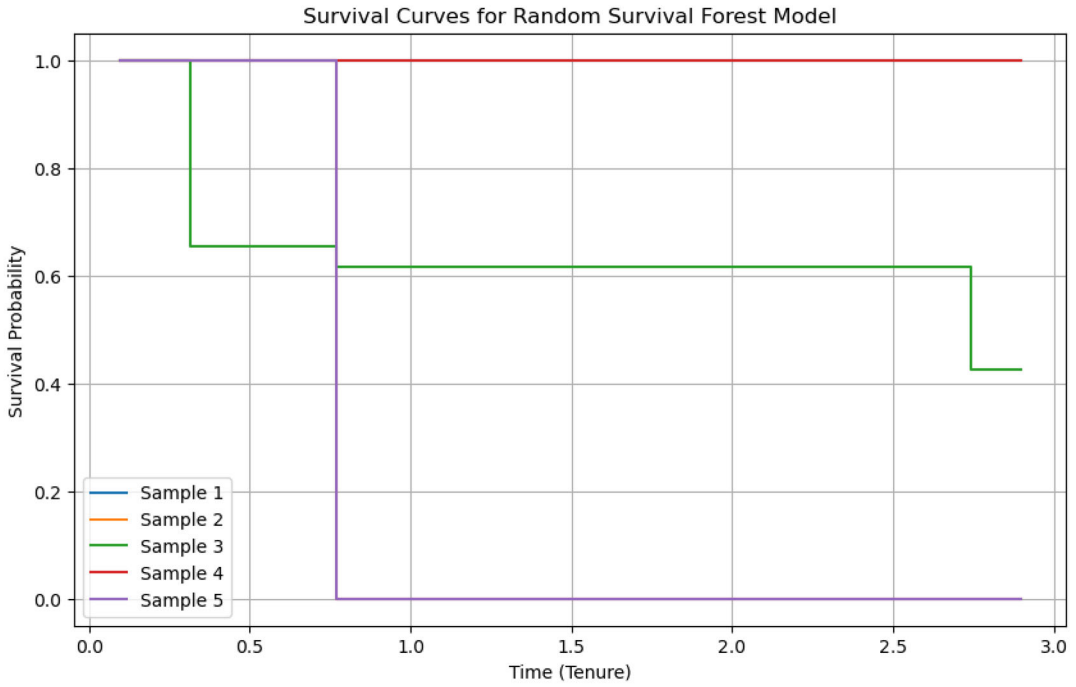
Each sample represents a group of employees or observations:

•**Sample 4 (Purple)**:
- Survival probability drops immediately to **0%**.
- **Interpretation**: This group experienced turnover very early, which could signal an issue such as poor onboarding, misaligned roles, or external factors leading to immediate exits.

•**Sample 3 (Green)**:
- Gradual decline in survival, stabilizing at around **45%** after 3 years.
- **Interpretation**: Some employees in this group remained employed long-term, indicating better retention compared to Sample 4.

•**Sample 5 (Red)**:
- Survival probability remains at **100%** throughout the observed tenure.
- **Interpretation**: This group experienced **no turnover**, possibly due to unique characteristics such as role type, better support systems, or external factors.

•**Samples 1 and 2 (Blue, Orange)**:
- Both maintain **100% survival** over the observed time, similar to Sample 5.
- **Interpretation**: These groups also experienced no turnover within the observation period, suggesting higher stability.



What to Look for Next

To gain more actionable insights:

•Stratify survival curves by **specific features** (e.g., delivered_packages, gender, or performance metrics).

•Compare survival trends across different **segments of the workforce**.

•Investigate **common characteristics** of groups like Sample 4 (high turnover) and Sample 5 (high retention).

Insights:

1.Kaplan-Meier:
- •Provides a basic, non-parametric survival estimate.
- •Moderate performance (C-Index = 0.69), useful as a baseline comparison.

2.Cox Proportional Hazards Model:
- •**Best-performing model** (C-Index = 0.72).
- •It effectively captures the relationship between covariates (e.g., delivered_packages, gender) and survival probabilities.
- •Strength: **Interpretability** – clear coefficients and hazard ratios explain the impact of predictors.

3.Random Survival Forest (RSF):
- •Performs poorly (C-Index = 0.5049), indicating weak predictive ability.
- •Likely due to **overfitting**, lack of sufficient data, or imbalance in features.
- •Strength: Can capture **nonlinear relationships**, but it struggles here compared to the Cox model.

Model Performance Summary

Model	C-Index	Interpretation
Kaplan-Meier	0.6900	Moderate predictive performance. A baseline, non-parametric approach.
Cox Proportional Hazards	0.7200	Best performance among the three models, showing strong predictive accuracy.
Random Survival Forest	0.5049	Poor predictive performance, close to random chance (C-Index ~ 0.5).

Key Takeaways:
- •The **Cox Proportional Hazards Model** outperforms the others and provides actionable insights into predictors of termination.
- •**Kaplan-Meier** serves as a useful baseline but lacks covariate adjustment.
- •**RSF** underperforms in this case, suggesting it may not be the right model for the current dataset or problem.

Next Steps:
1.Focus on refining and interpreting the **Cox model** further.
2.Investigate why RSF underperformed (e.g., data quality, feature importance).
3.Use the insights gained from the Cox model (e.g., delivered_packages and gender) to drive retention strategies.

Summary of Findings

The **Cox Proportional Hazards Model** was the best-performing model with a Concordance Index (C-Index) of **0.7200**, demonstrating strong predictive accuracy and clear interpretability. The **Kaplan-Meier Estimator** achieved moderate performance with a C-Index of **0.6900**, serving as a solid baseline model. In contrast, the **Random Survival Forest (RSF)** performed poorly with a C-Index of **0.5049**, indicating limited predictive power and possible overfitting.

From the Cox model, **delivered packages** emerged as the strongest predictor of employee retention. Higher delivery volumes were associated with a significantly lower risk of turnover. **Shipments per on-zone hour** had a smaller, moderate impact, while gender-related features showed mixed results. In the Cox model, males demonstrated a lower risk of turnover; however, in the RSF model, gender features carried no importance. Results for employees who chose not to self-identify were unreliable due to sparse data.
The **Kaplan-Meier Survival Curves** revealed that turnover risk is highest within the first **1.5 years** of employment. Significant attrition occurs during this period, emphasizing the need for targeted early retention strategies. Survival patterns from the **Random Survival Forest** varied across samples, with some groups experiencing immediate attrition and others showing complete retention throughout the observed tenure.

Overall, **delivered packages** is the most influential factor for predicting retention, while turnover tends to peak in the early months of employment. The **Cox Proportional Hazards Model** offers the most reliable insights into the drivers of turnover, making it the preferred tool for informing retention strategies.

To improve retention, focus on incentivizing and supporting high-performing employees while addressing early turnover through improved onboarding and engagement programs. Data issues, such as sparse representation of certain groups, should also be addressed to enhance future modeling efforts.