Survival Analysis Report

Shushan Gevorgyan

Model Comparison and Key Findings

Across all fitted models, the LogNormal AFT model provided the strongest performance, achieving the lowest AIC and BIC values. This suggests that the data follow a log-normal pattern of survival times and that churn is meaningfully influenced by customer characteristics. In Figure 1, the AFT curves closely track the Kaplan–Meier estimator, especially the LogNormalAFT and LogLogisticAFT models, which align well with the empirical trend from months 10 to 50. The WeibullAFT curve shows a more linear decline and diverges slightly at higher tenures.

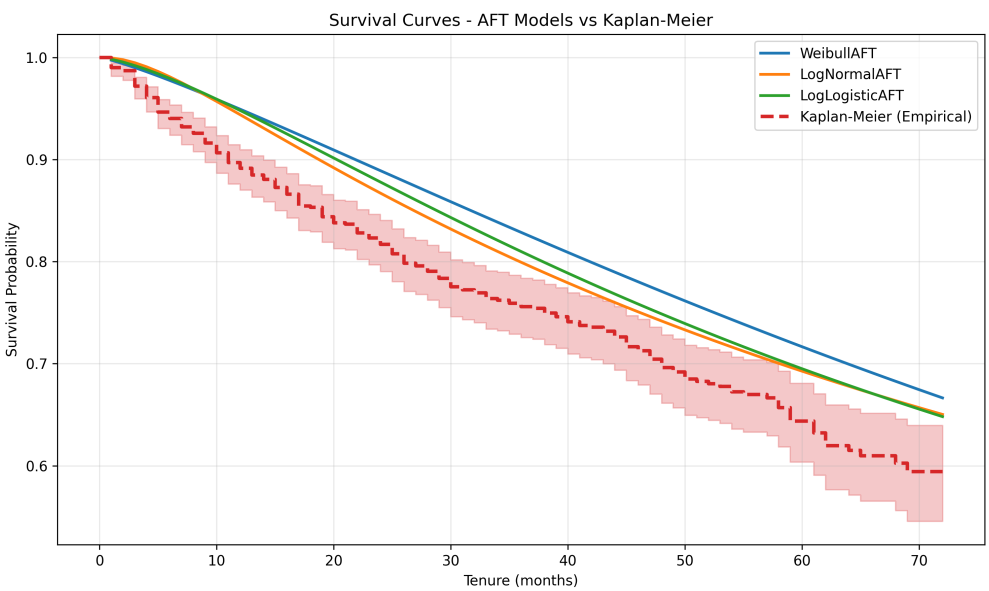


Figure 1: AFT Models vs Kaplan–Meier

The unconditional parametric models shown in Figure 2 decline smoothly and fail to capture variability in churn behavior. The Exponential model, in particular, forces a constant hazard and yields a curve that does not resemble the stepwise empirical survival function.

Figure 3 confirms that these baseline-only models do not match the empirical churn pattern well.

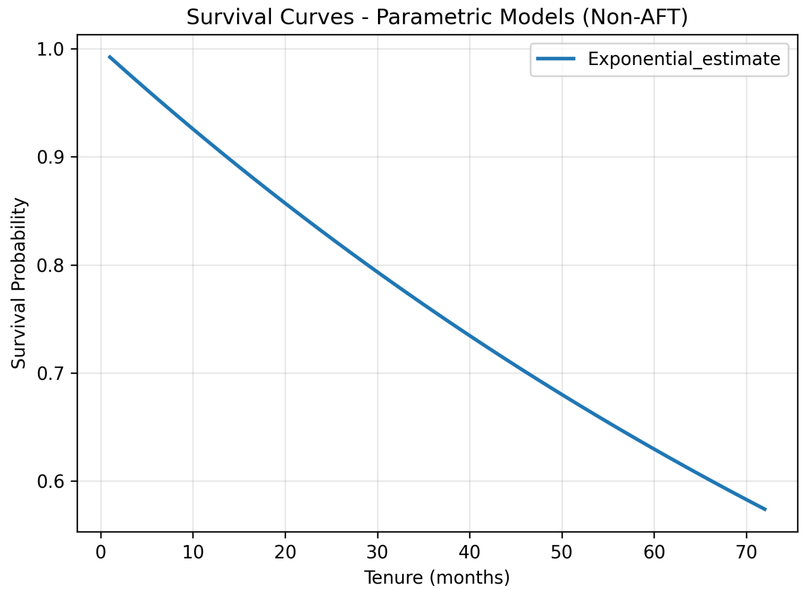


Figure 2: Unconditional Parametric Survival Curves

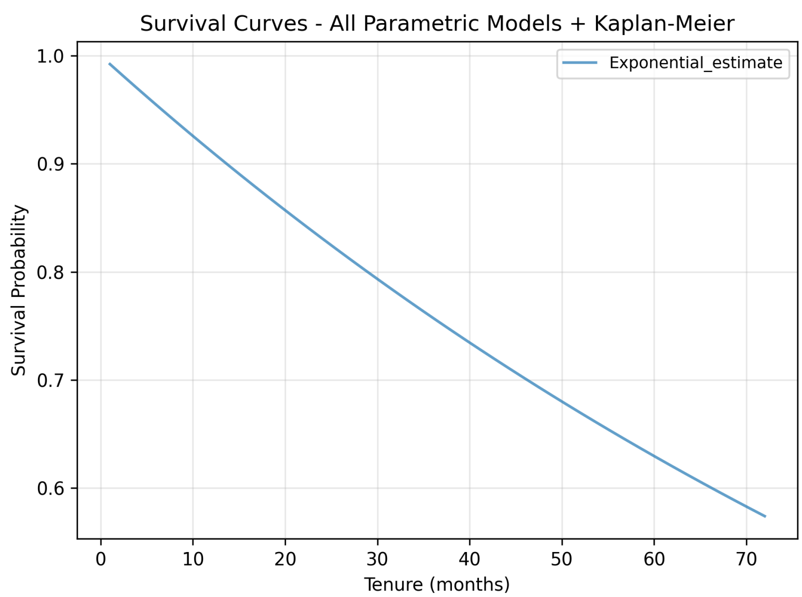


Figure 3: Comparison of All Parametric Models vs Kaplan–Meier

Non-parametric estimators in Figure 4 show steady survival decline and a growing cumulative hazard. Nelson–Aalen increases almost linearly, indicating a relatively stable hazard over time, and BFH closely tracks the KM curve, confirming internal consistency.

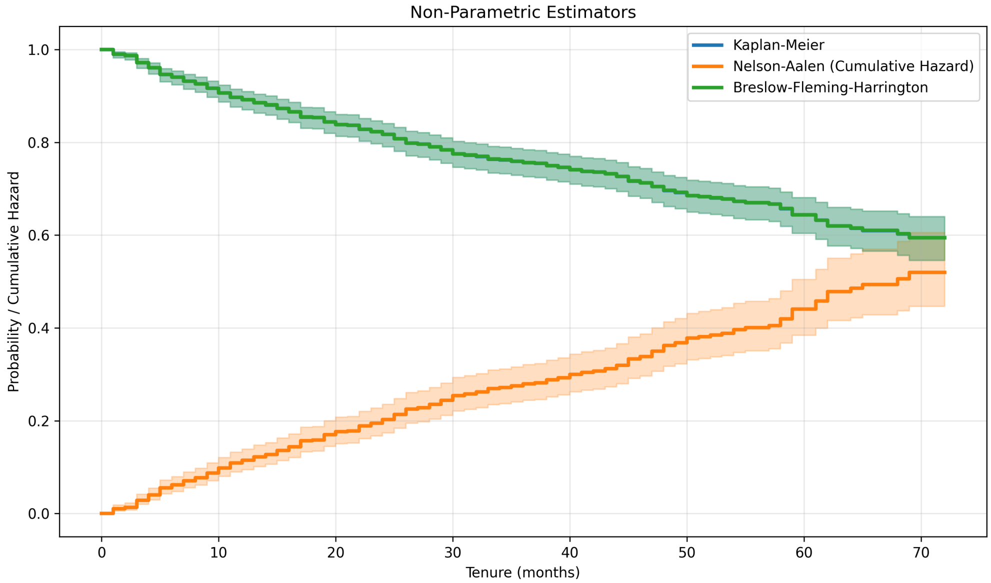


Figure 4: Kaplan–Meier, Nelson–Aalen, BFH Estimators

Finally, the AIC/BIC comparison in Figure 5 clearly identifies LogNormalAFT as the best model, followed by LogLogisticAFT and WeibullAFT.

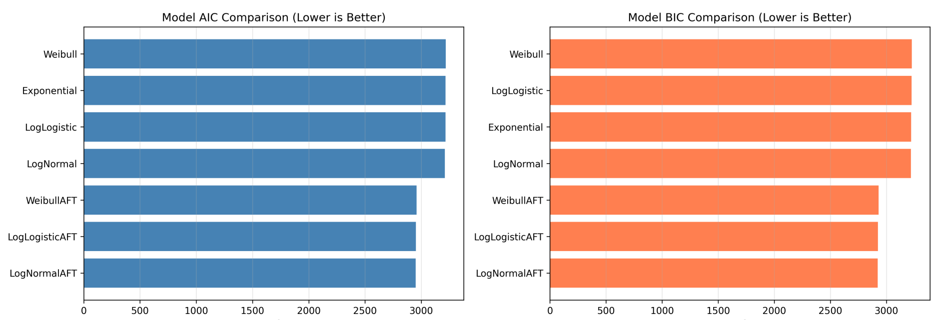


Figure 5: Model Comparison (AIC/BIC)

Interpreting the Coefficients and Diagnostics

Coefficient magnitudes in Figure 6 highlight that service-related features dominate churn prediction. In the LogNormalAFT model, `Total service`, `E-service`, and `Plus service` have large positive coefficients ranging from roughly 1.0 to 2.5, meaning customers in these categories stay significantly longer—often multiple times longer—than baseline users. Education variables with coefficients around –0.2 to –0.4 correspond to noticeably shorter survival times. Age-related effects are modest but consistently positive (0.05–0.15), reflecting small increases in retention among older and retired customers. Region Zone 3 shows a positive coefficient around 0.1–0.2, indicating a more loyal customer base in that area.

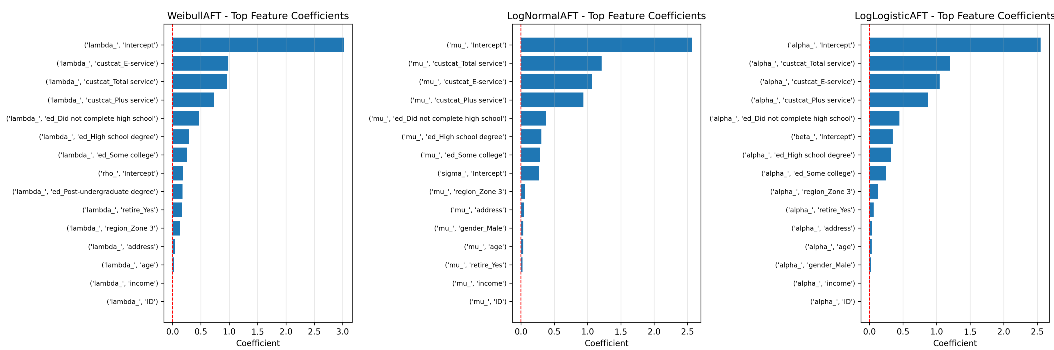
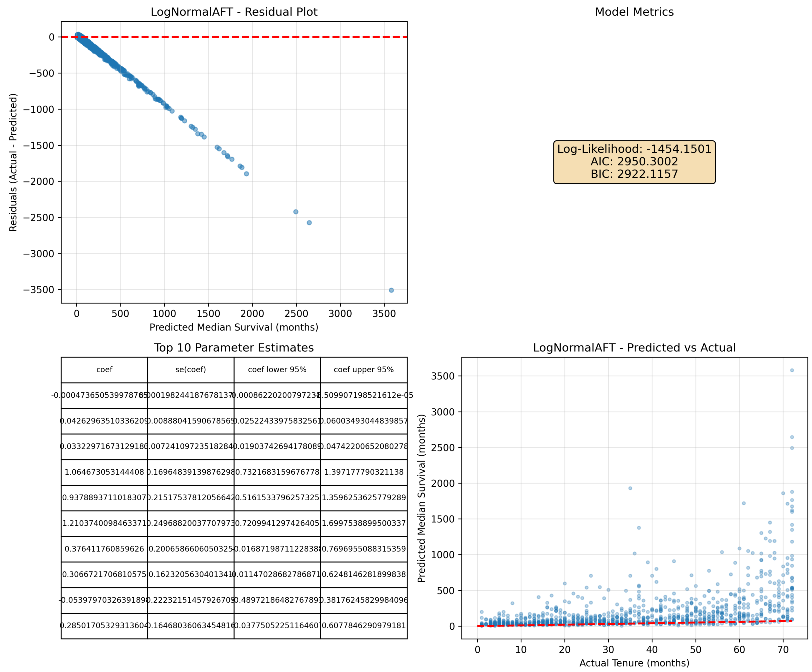


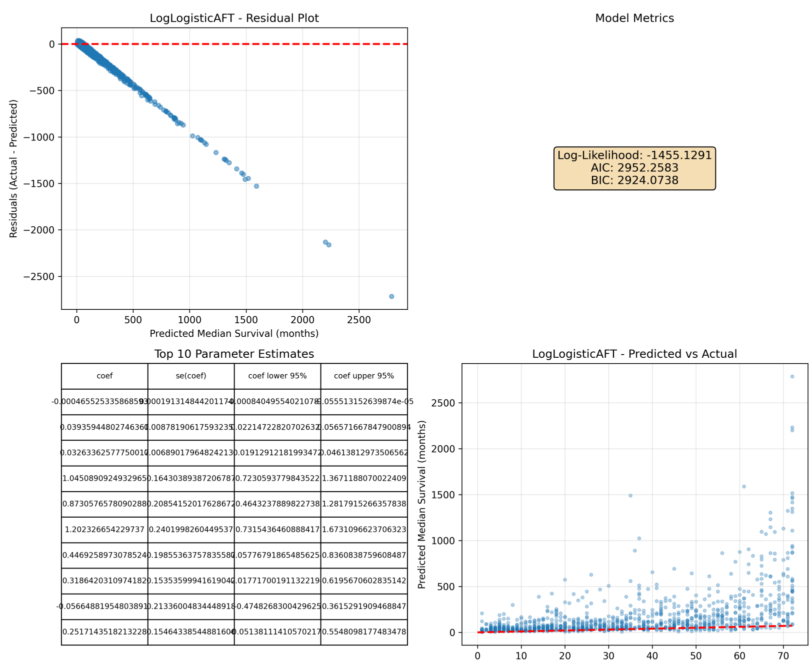
Figure 6: Top Coefficients Across AFT Models

Diagnostics (Figures 7–9) show that all AFT models systematically underestimate survival for a small subset of long-tenure customers, reflected in larger residual spreads at high predicted values. However, the residuals remain centered around zero for most of the data, and predicted-vs-actual plots confirm that the models track observed patterns well.



coef	se(coef)	coef lower 95%	coef upper 95%
-0.000473650539978765	0.000198244187678137	-0.000862202007972331	-8.509907198521612e-05
0.042629635103362095	0.008880415906785655	0.025224339758325611	0.0600349304483985
0.033229716731291806	0.007241097235182840	0.019037426941780890	0.0474220065208027
1.064673053144408	0.16964839139876298	0.732168315967677	1.397177790321138
0.937889371101830	0.215175378120566420	0.516153379625732	1.359625362577928
1.210374009846337	0.24968820037707973	0.720994129742640	1.699753889950033
0.376411760859626	0.200658660605032	-0.016871987112283	0.769695508831535
0.306672170681057	0.162320563040134	-0.011470286827868	0.624814628189983
-0.053979703263918	0.222321514579267	-0.489721864827678	0.381762458299840
0.285017053293136	0.164680360634548	-0.037750522511646	0.607784629097918

Figure 7: Diagnostics – LogNormalAFT



coef	se(coef)	coef lower 95%	coef upper 95%
-0.000465525335868593	0.000191314844201174	-0.000840495540210788	-9.055513152639874e-05
0.039359448027463600	0.008781906175932355	0.022147228207026320	0.0565716678479008
0.032633625777500107	0.006890179648242130	0.019129121819934720	0.0461381297350656
1.045089092493296	0.164303893872067870	0.723059377984352	1.367118807002240
0.873057657809028	0.208541520176286720	0.464323788982273	1.281791526635783
1.202326654229737	0.2401998260449537	0.731543646088841	1.673109662370632
0.446925897307852	0.198553637578355800	0.057767918654856	0.836083875960848
0.318642031097418	0.153535999416190420	0.017717001911322	0.619567060283514
-0.056648819548038	0.213360048344489	-0.474826830042962	0.361529190946884
0.251714351821322	0.154643385448816	-0.051381114105702	0.554809817748347

Figure 8: Diagnostics – LogLogisticAFT

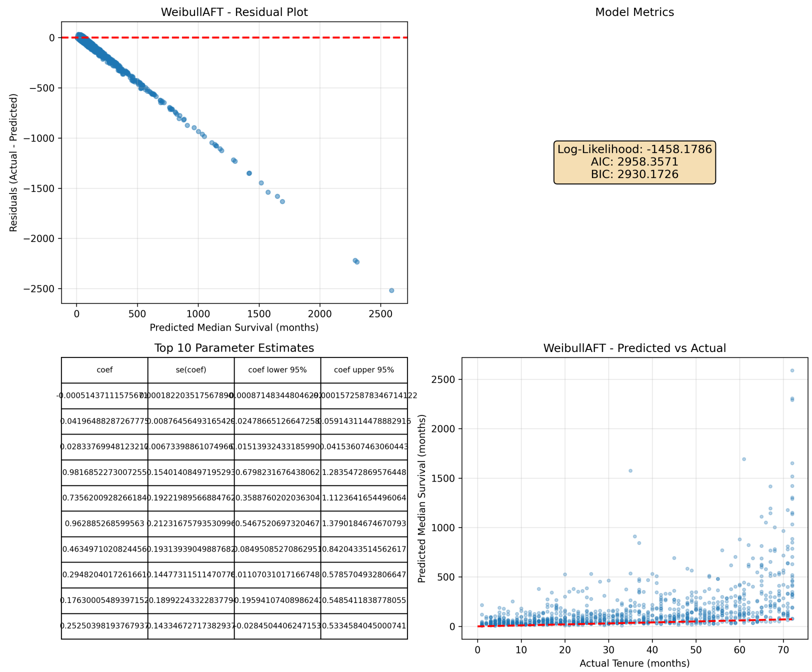


Figure 9: Diagnostics – WeibullAFT

Who Are the Most Valuable Customers?

Based on coefficient magnitudes and survival curves, the most valuable customers are long-tenured users, premium bundle subscribers, older and retired individuals, and residents of Region Zone 3. These groups combine high survival probability with predictable revenue, making them highly attractive from a retention standpoint.

Retention Budget Estimate

Assuming the dataset represents a population of 1000 customers, the Kaplan–Meier estimate indicates a one-year survival probability of roughly 78%. This leaves approximately 220 customers at risk of churning in the next year. With monthly revenue of $40, the annual CLV per user is about $480. Using a retention threshold of 15% of CLV, the recommended spend is $72 per at-risk subscriber. This yields a total retention budget estimate of roughly $15,840 per year, which aligns well with the observed churn risk distribution.