Build a churn model with the following data

Jonathan P. Latner, PhD

November 7, 2022

Overview

1. Introduction
2. Summary statistics
3. GLM
4. Compare
5. Survival
6. Segmentation
7. Conclusion

Variables

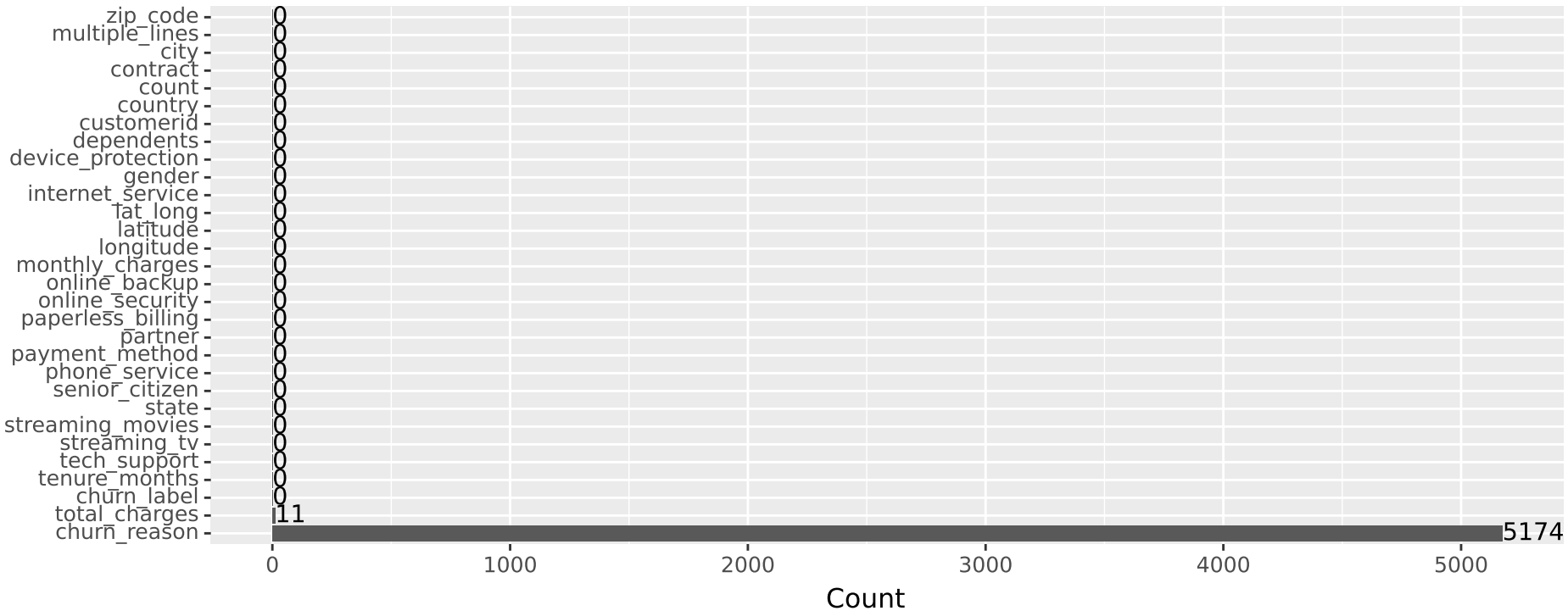
Table 1:

number	variable	type	count
1	state	factor	1
2	country	factor	1
3	churn_label	factor	2
4	senior_citizen	factor	2
5	phone_service	factor	2
6	dependents	factor	2
7	partner	factor	2
8	gender	factor	2
9	paperless_billing	factor	2
10	streaming_movies	factor	3
11	online_security	factor	3
12	online_backup	factor	3
13	multiple_lines	factor	3
14	internet_service	factor	3
15	device_protection	factor	3
16	contract	factor	3
17	streaming_tv	factor	3
18	tech_support	factor	3
19	payment_method	factor	4
20	churn_reason	factor	21
21	city	factor	1129
22	lat_long	factor	1652
23	customerid	factor	7043
24	zip_code	numeric	1
25	latitude	numeric	1
26	longitude	numeric	1
27	tenure_months	numeric	1
28	monthly_charges	numeric	1
29	total_charges	numeric	1

Missing values

Who are the missings *total_charges*? Customers in their first month

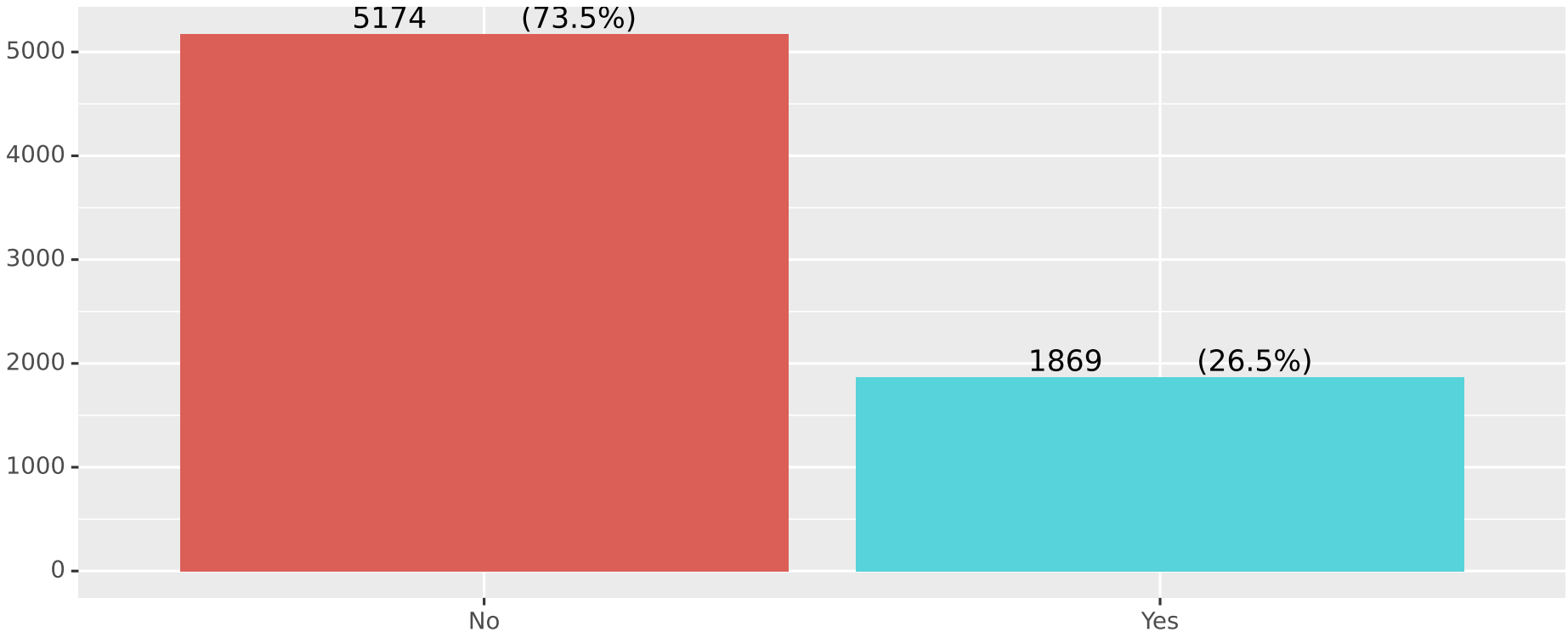
Figure 1:



Customer churn

Need to over sample the training dataset because only 27% churn (DV)

Figure 2:



Customer churn by demographics

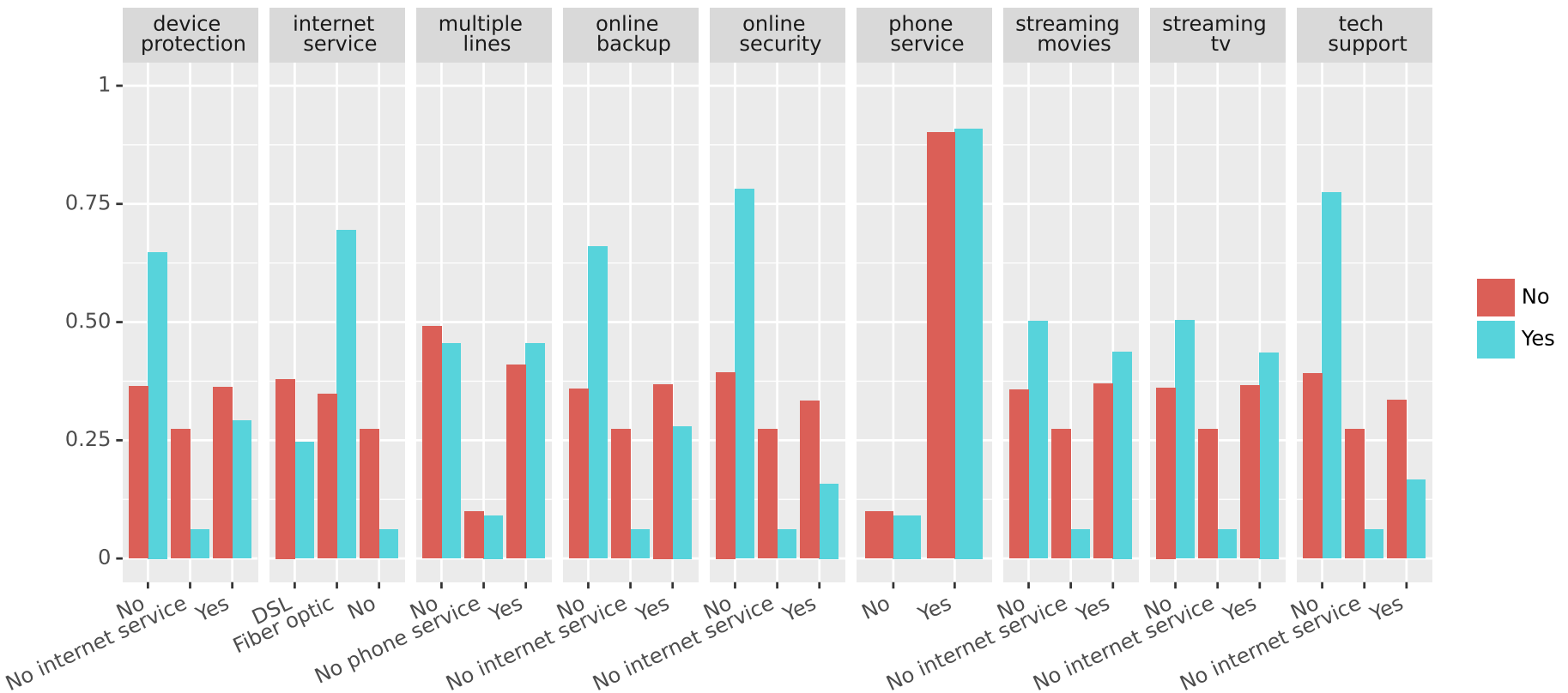
Households w/o children, single, and seniors more likely to churn

Figure 3:

Customer churn by services

Households w/ internet, but w/o online services (except streaming) most likely to churn

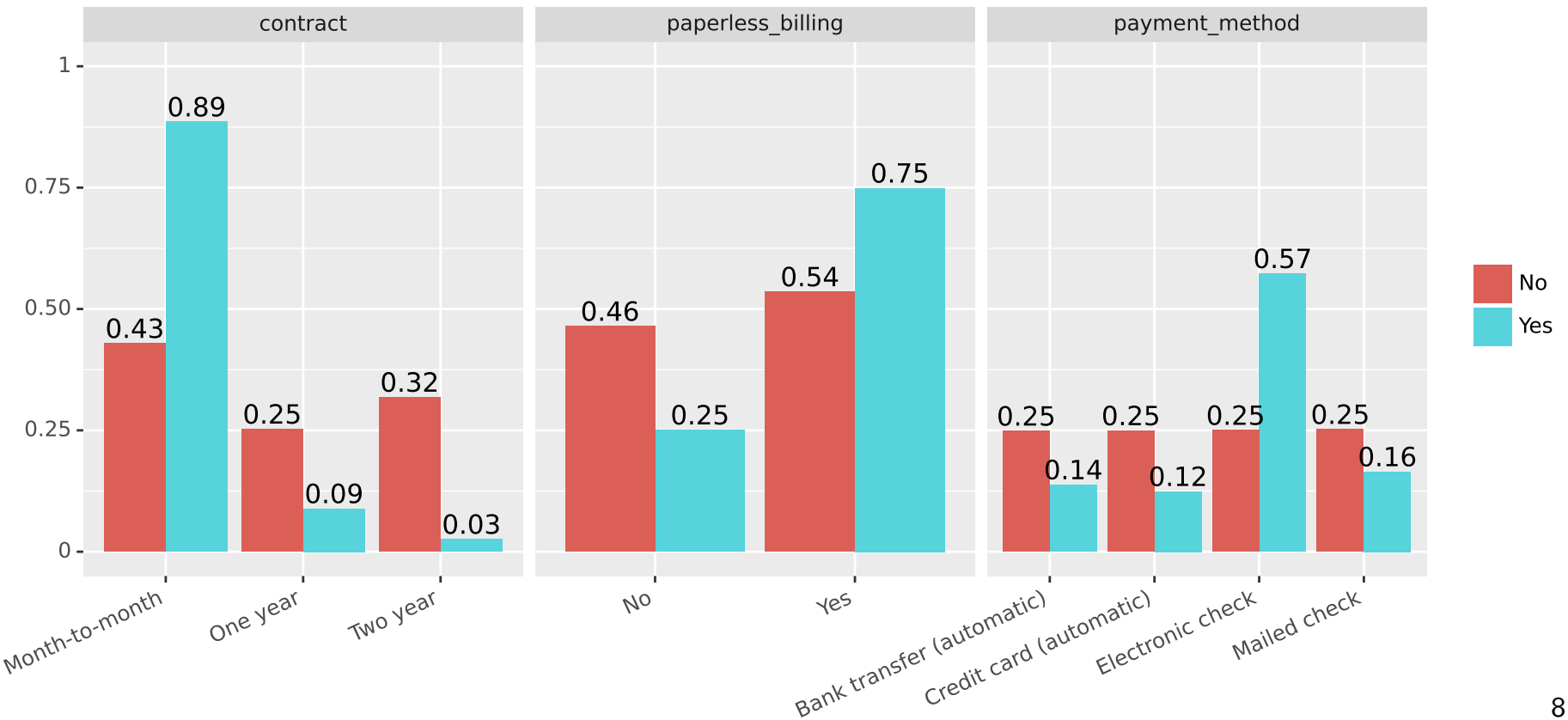
Figure 4:



Customer churn by contract type

Households w/ monthly contracts, paperless billing (i.e. contact), or electronic checks most likely to churn

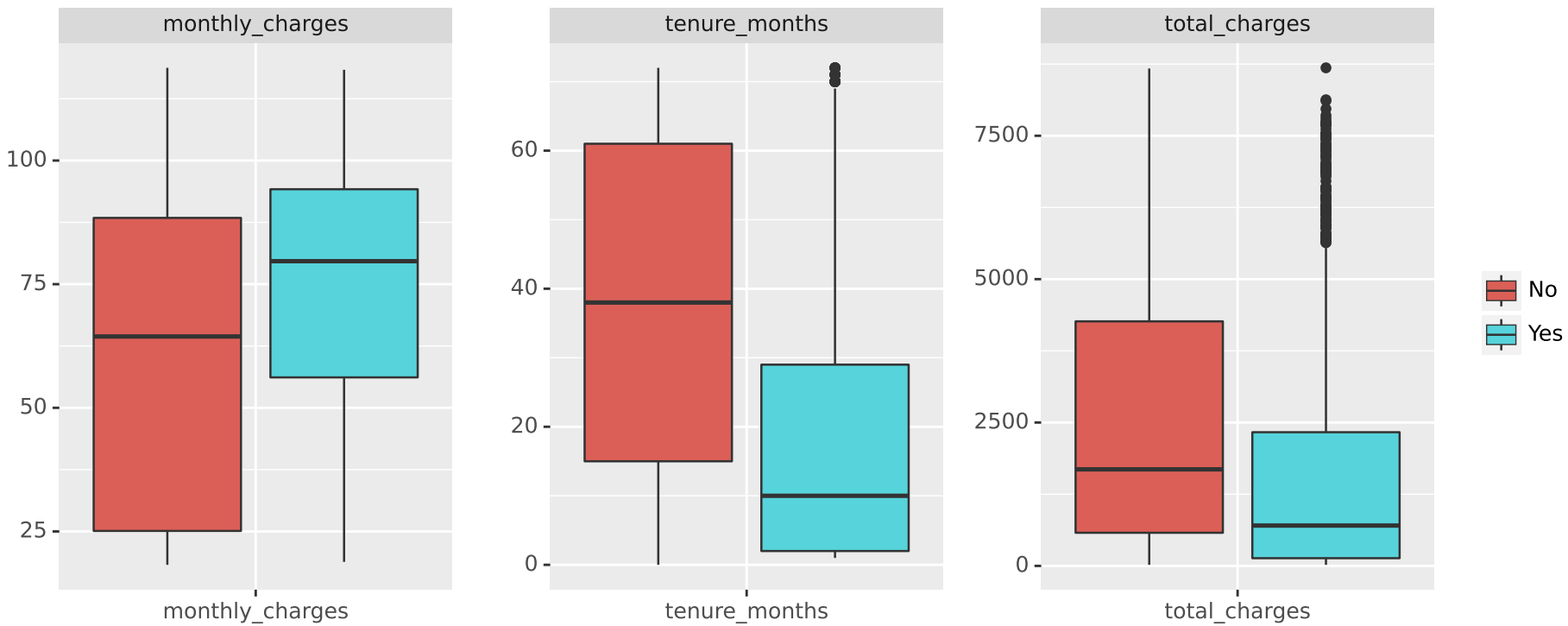
Figure 5:



Customer churn by contract charges

Higher monthly charges, lower tenure months, and higher total charges more likely to churn.
Tenure explains inverse relationship between total charges and churn.

Figure 6:

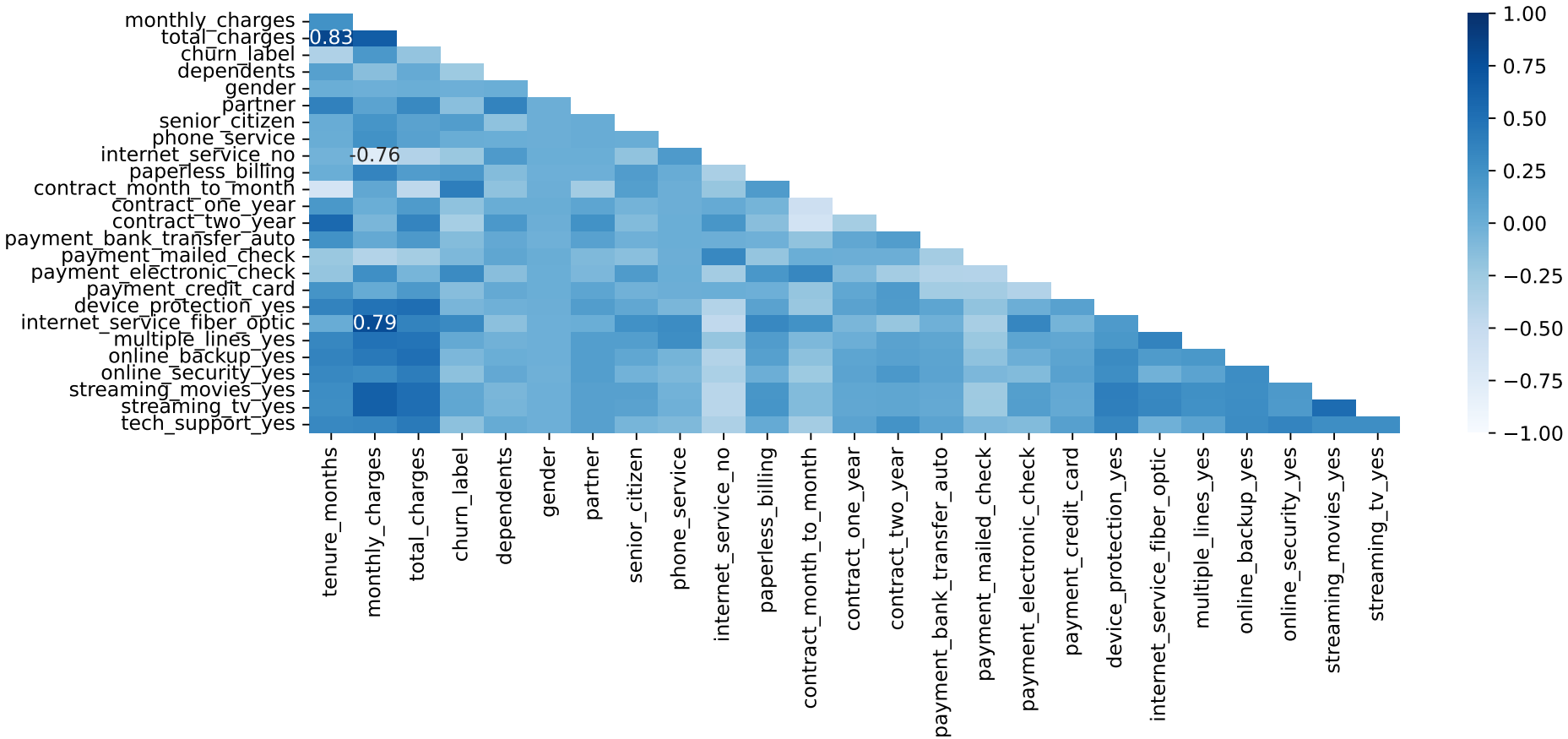


Multivariate correlation

High correlation between *tenure_months*, *monthly_charges*, *total_charges*

High correlation between *monthly_charges*, *internet_service_no*, *internet_service_fiber_optic*

Figure 7:



Variance inflation factors (VIF)

Drop *monthly_charges*, *total_charges*, *phone_service*?

Keep *tenure_months*?

Table 2:

Features	VIF_1	VIF_2	VIF_3	VIF_4
monthly_charges	13.70	NaN	NaN	NaN
total_charges	10.88	9.63	NaN	NaN
phone_service	10.51	10.33	10.33	NaN
internet_service_fiber_optic	9.34	4.99	4.45	3.47
tenure_months	9.28	9.27	2.51	2.07
internet_service_no	5.44	2.82	2.66	1.92
payment_electronic_check	3.50	3.20	2.99	2.90
paperless_billing	3.30	3.11	3.08	2.99
streaming_tv_yes	3.07	2.74	2.67	2.67
streaming_movies_yes	3.06	2.74	2.67	2.67
contract_two_year	3.05	2.98	2.86	2.79
multiple_lines_yes	2.60	2.48	2.44	2.27
partner	2.28	2.24	2.23	2.21
device_protection_yes	2.08	2.07	2.03	2.02
payment_bank_transfer_auto	2.05	1.91	1.83	1.76
payment_credit_card	1.99	1.88	1.82	1.75
tech_support_yes	1.97	1.97	1.92	1.88
gender	1.97	1.87	1.87	1.82
contract_one_year	1.92	1.87	1.85	1.81
online_backup_yes	1.89	1.89	1.84	1.82
online_security_yes	1.76	1.76	1.74	1.67
dependents	1.48	1.48	1.47	1.47
senior_citizen	1.33	1.32	1.32	1.32

Accuracy

- No real gain/loss in accuracy with different GLM model specifications
- Choose model 4: Drop *monthly_charges*, *total_charges*, *phone_service*
  - total charges are a result of monthly charges and tenure
  - Monthly charges are a result of services.
  - 90% of customers have phone service & no difference in churn

Table 3:

measure	model_1	model_2	model_3	model_4
TN	0.390	0.377	0.385	0.385
FP	0.110	0.123	0.115	0.115
FN	0.066	0.067	0.081	0.082
TP	0.434	0.433	0.419	0.418
ACC	0.824	0.810	0.804	0.803

Compare training models on test data

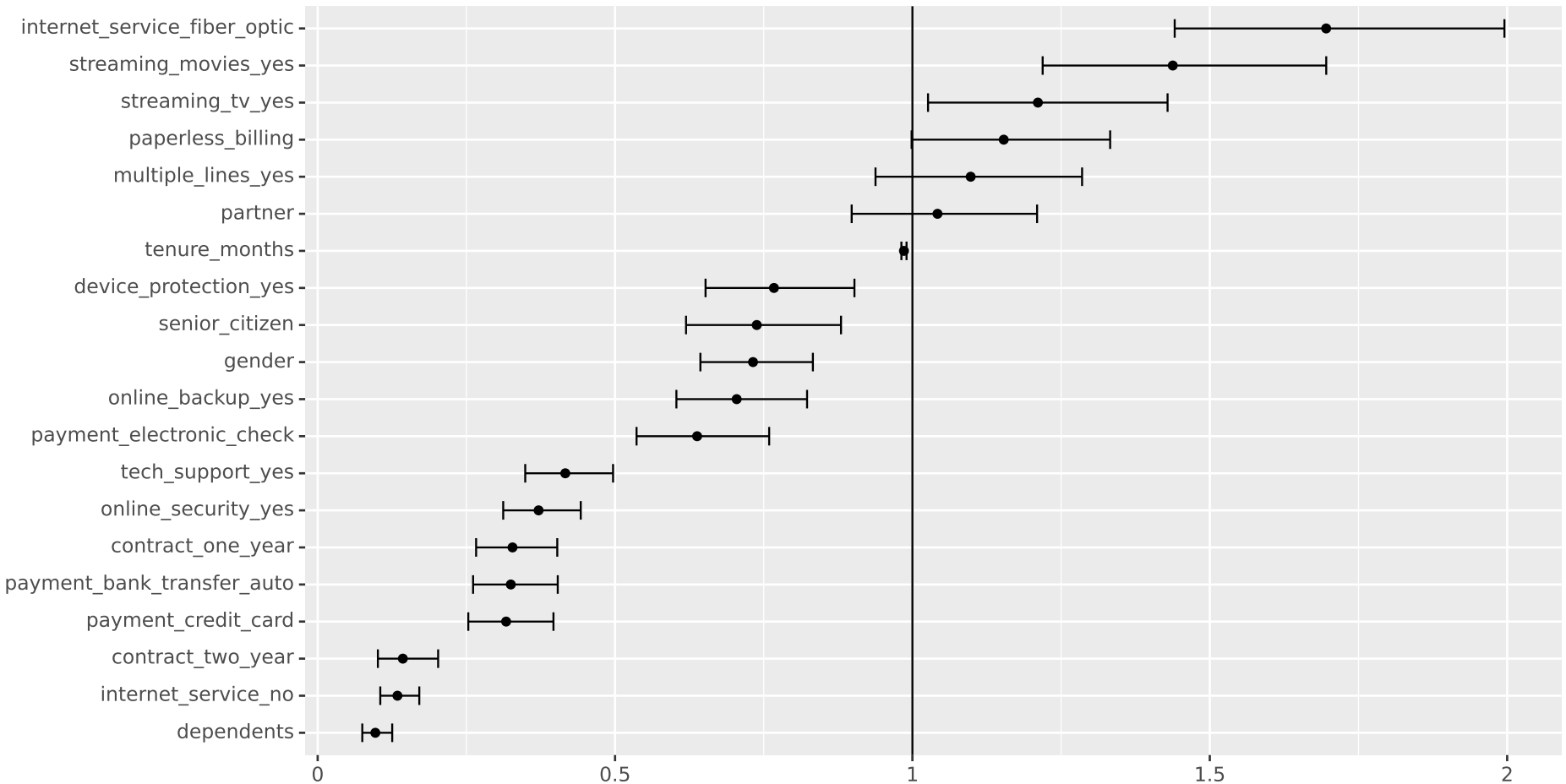
Random forest (RF) is most accurate, but GLM higher true positive (TP)
Choose GLM model

Table 4:

measure	GLM	KNN	NB	DT	RF
TN	0.557	0.490	0.563	0.574	0.592
FP	0.174	0.242	0.169	0.157	0.140
FN	0.072	0.045	0.073	0.111	0.100
TP	0.196	0.224	0.196	0.157	0.168
ACC	0.754	0.714	0.759	0.731	0.760

Examine preferred GLM model

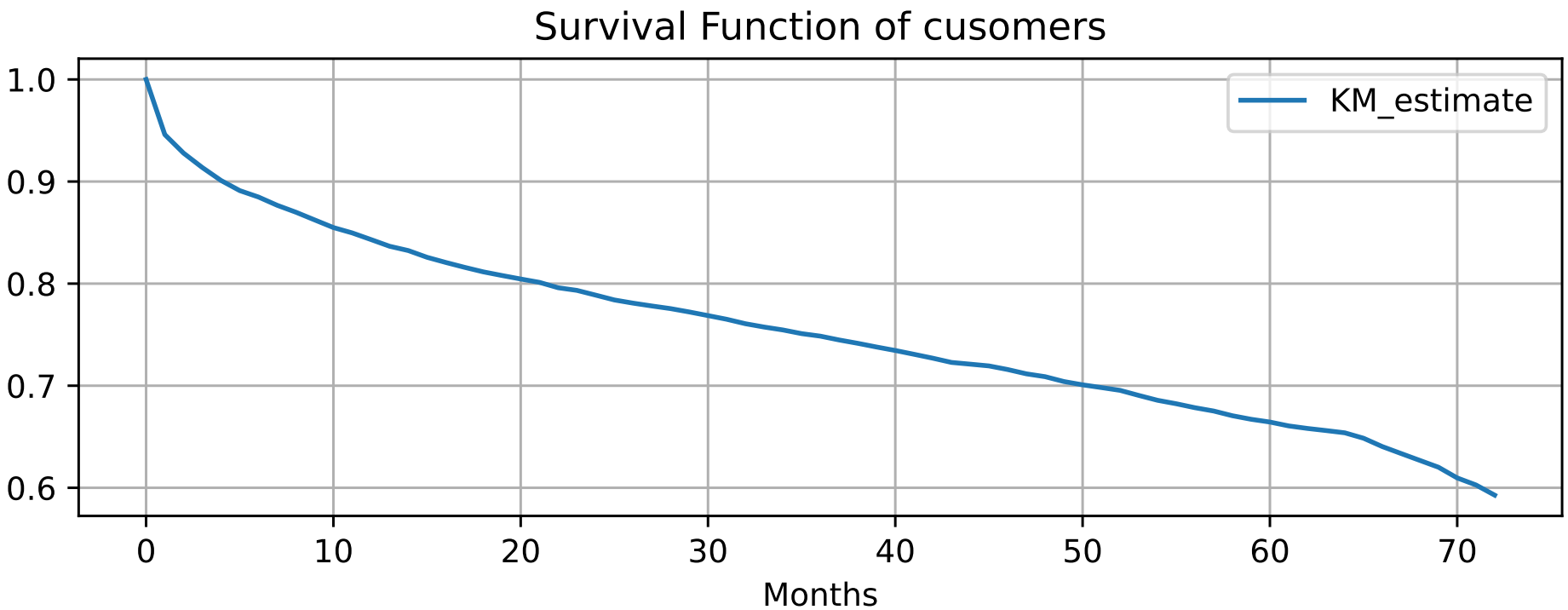
Figure 8: Odds ratios



KM curve

After 72 months (the max tenure in our data), the company can retain ≈ 60% of its customers.

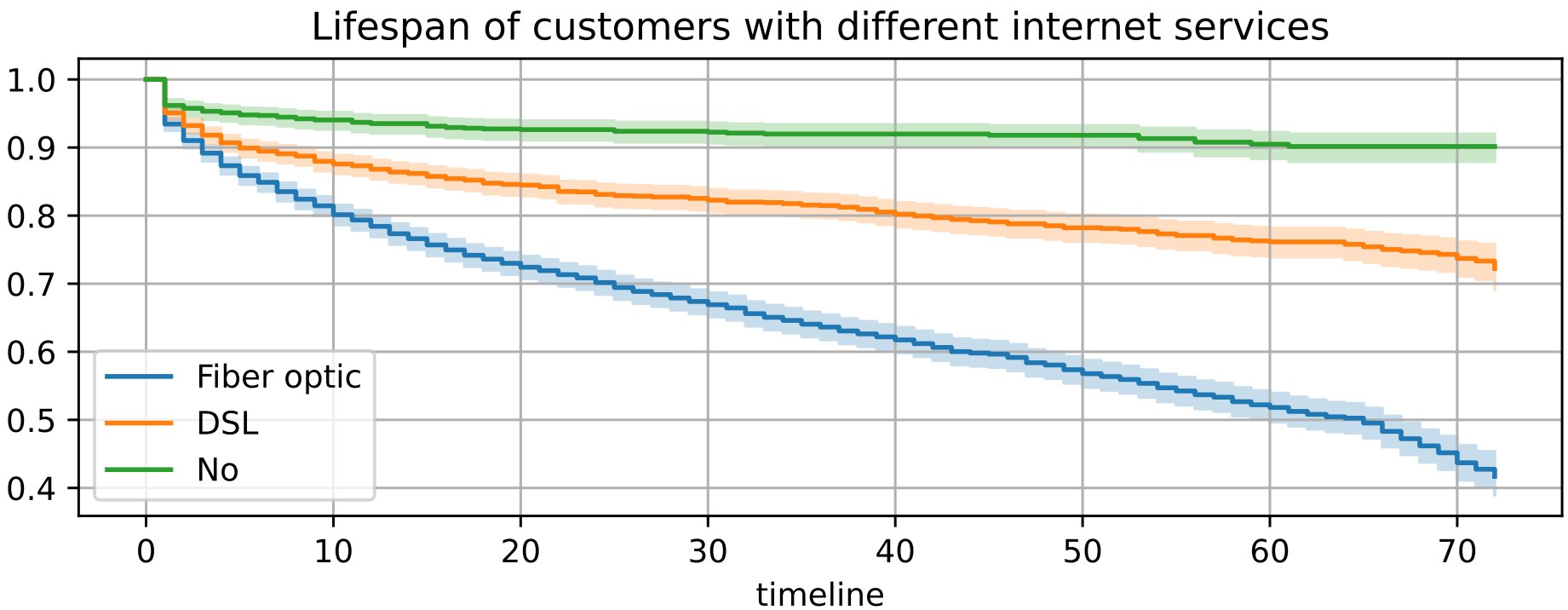
Figure 9:



KM curve by internet service

After 72 months, the company retains 42% Fiber optic, 72% DSL, and 90% without internet. Is internet not good or too expensive?

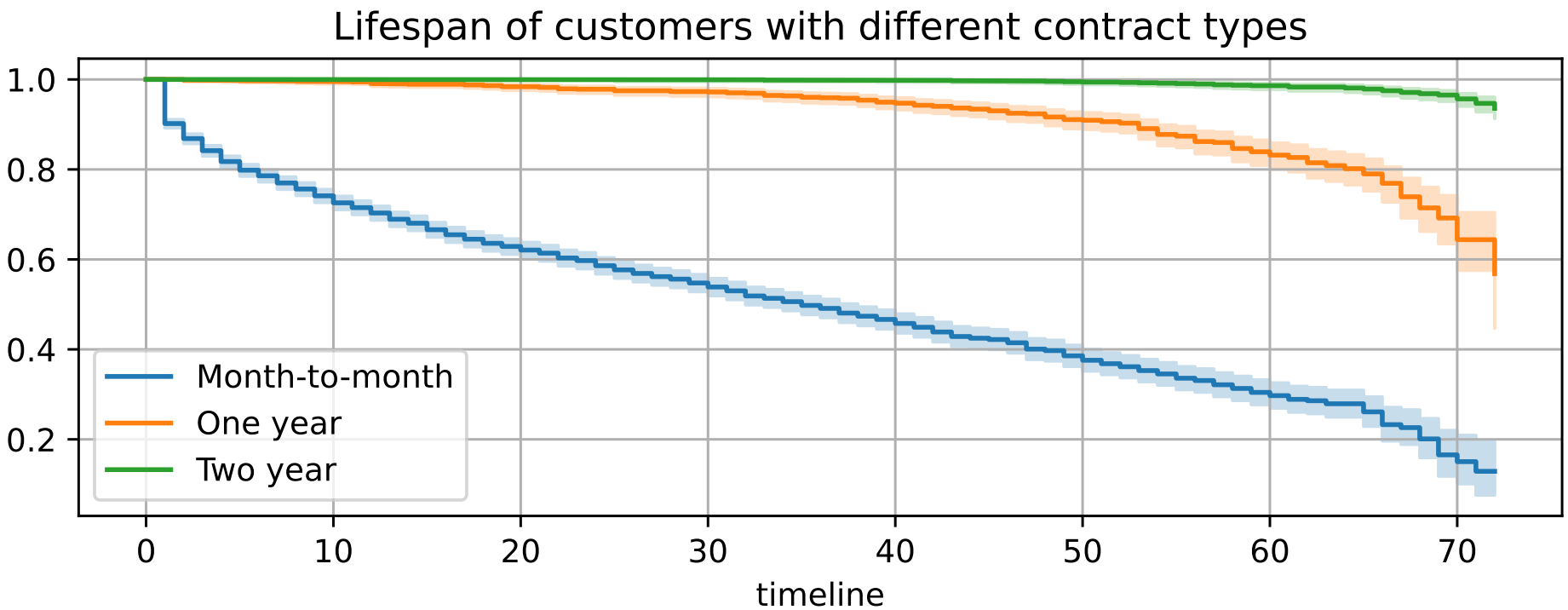
Figure 10:



KM curve by contract type

After 72 months, the company retains 13% month-to-month, 57% one year contract, and 94% two year contract

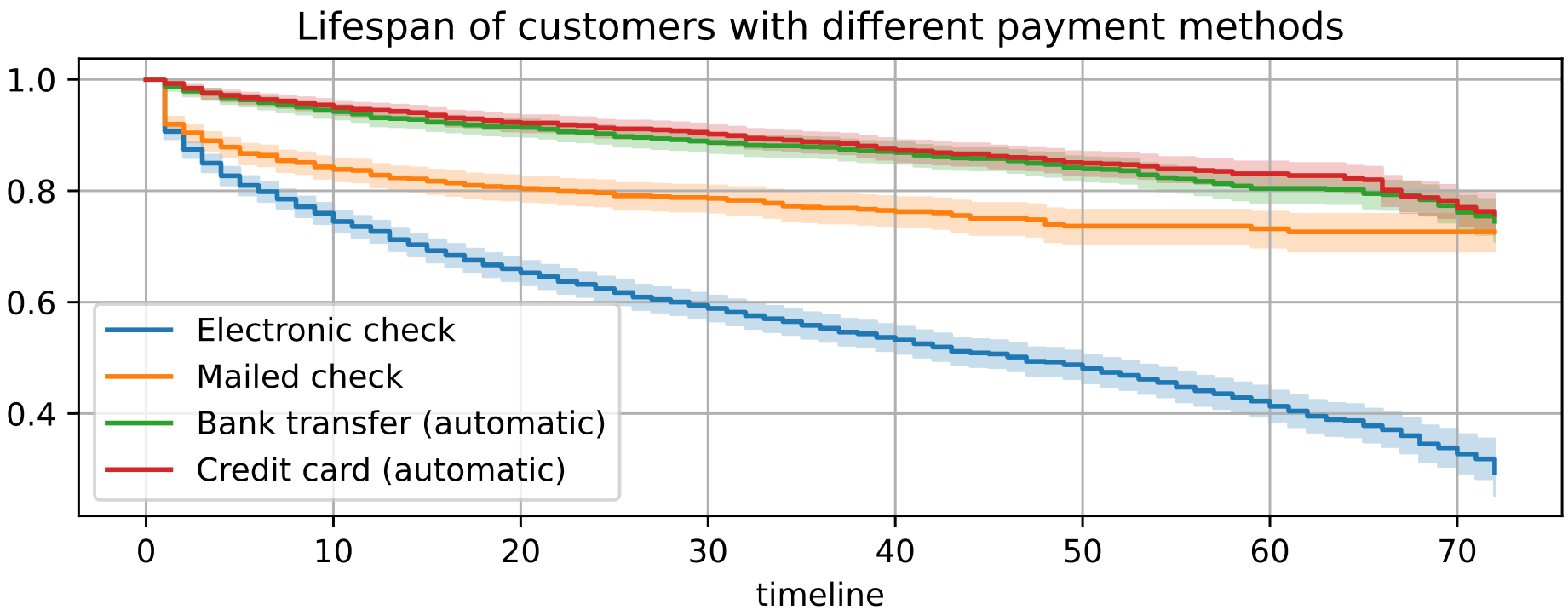
Figure 11:



KM curve by payment type

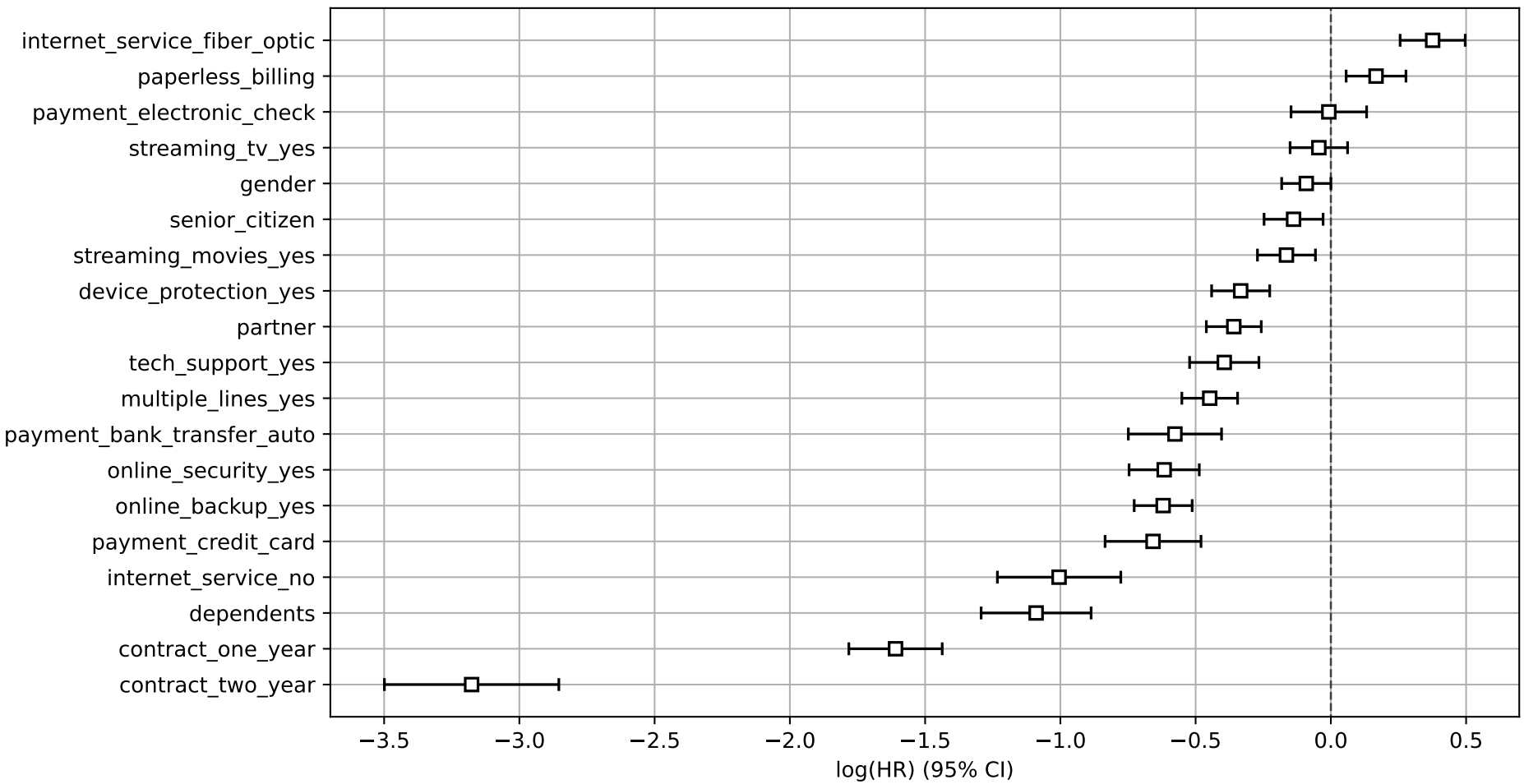
After 72 months, the company retains 29% electronic check compared to $\approx$ 75% other forms of payment

Figure 12:



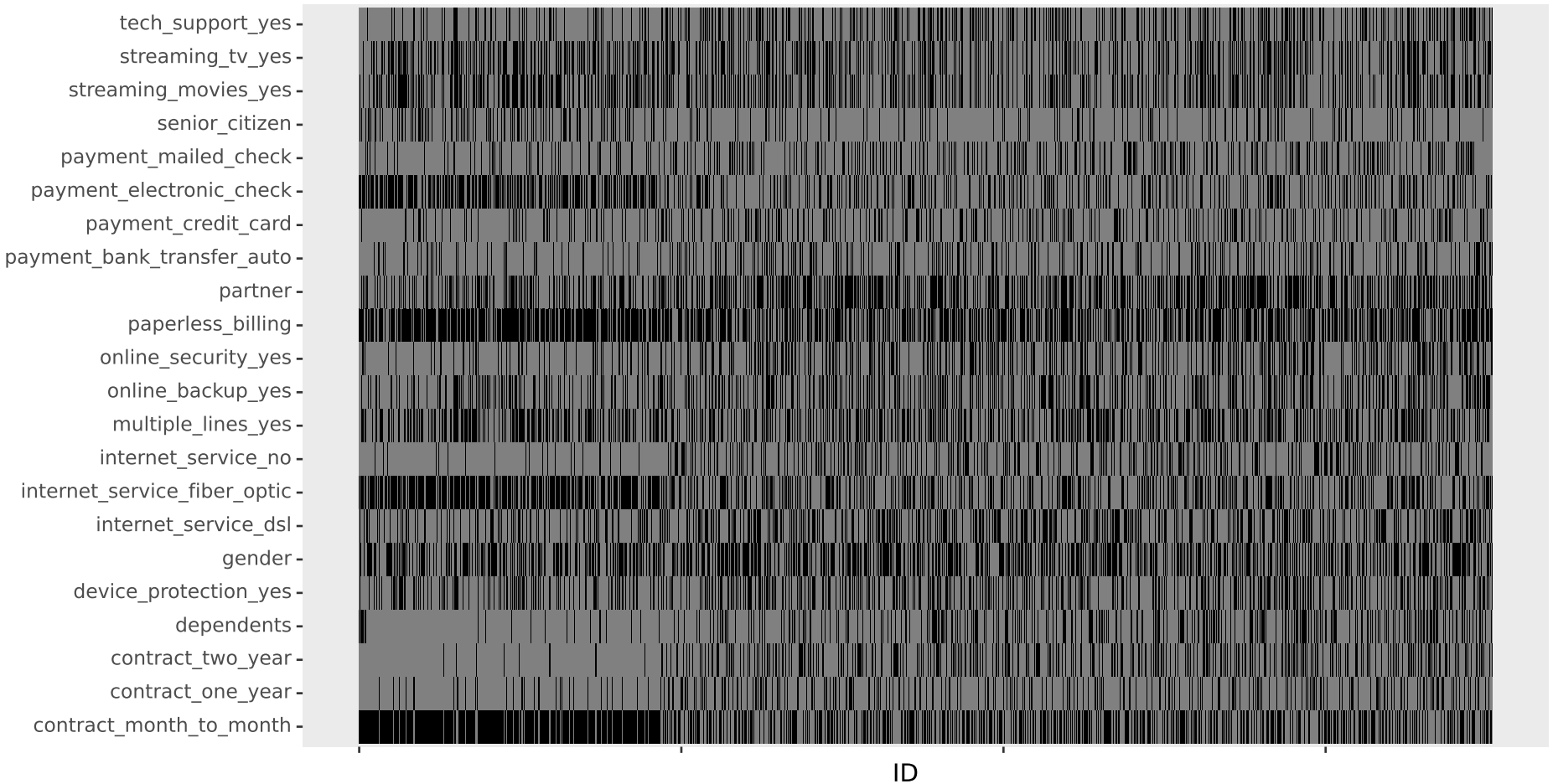
Cox proportional hazard (CPH) model

Figure 13:



Customers by all available services

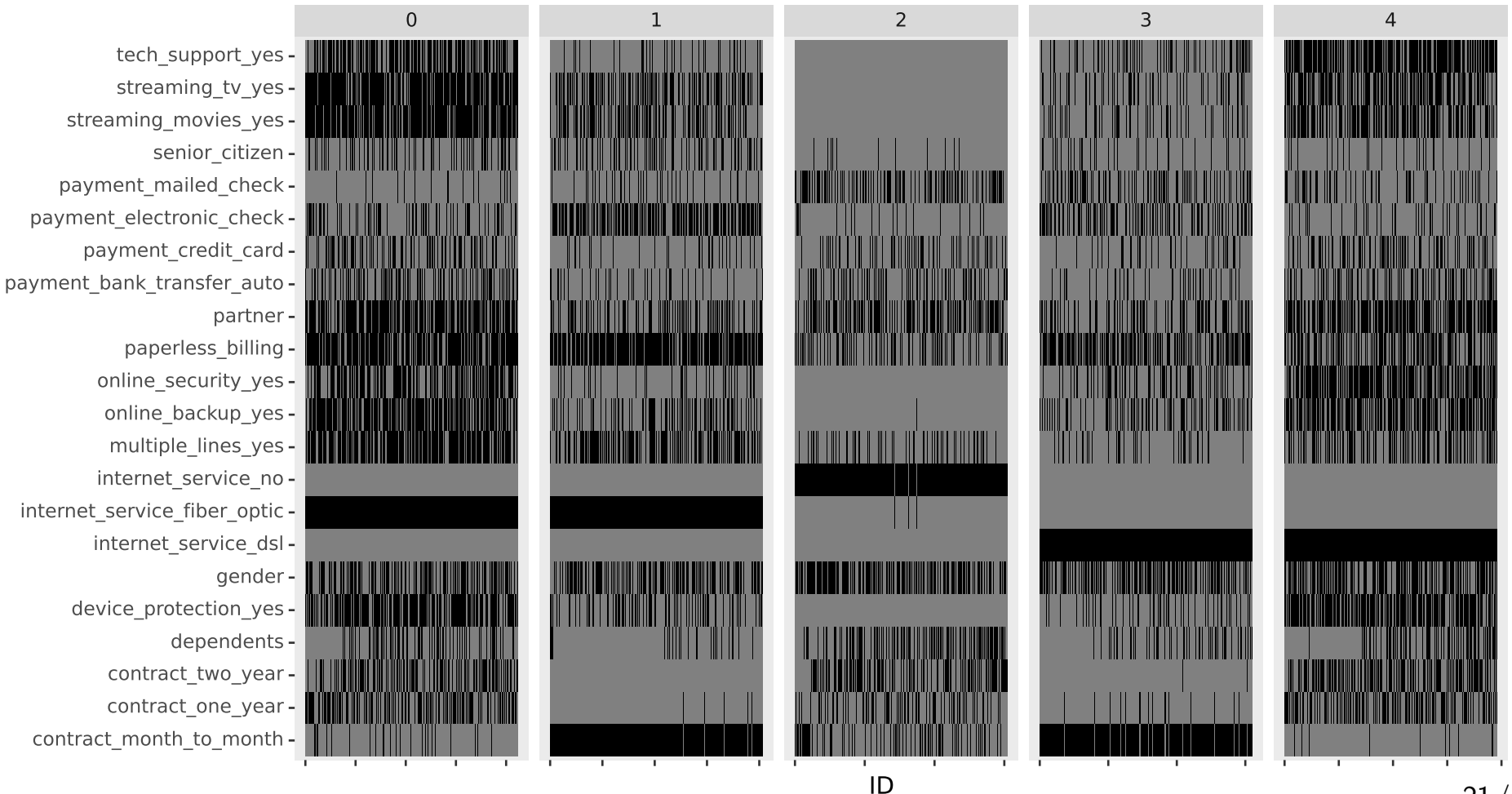
Figure 14:



Customers by segment

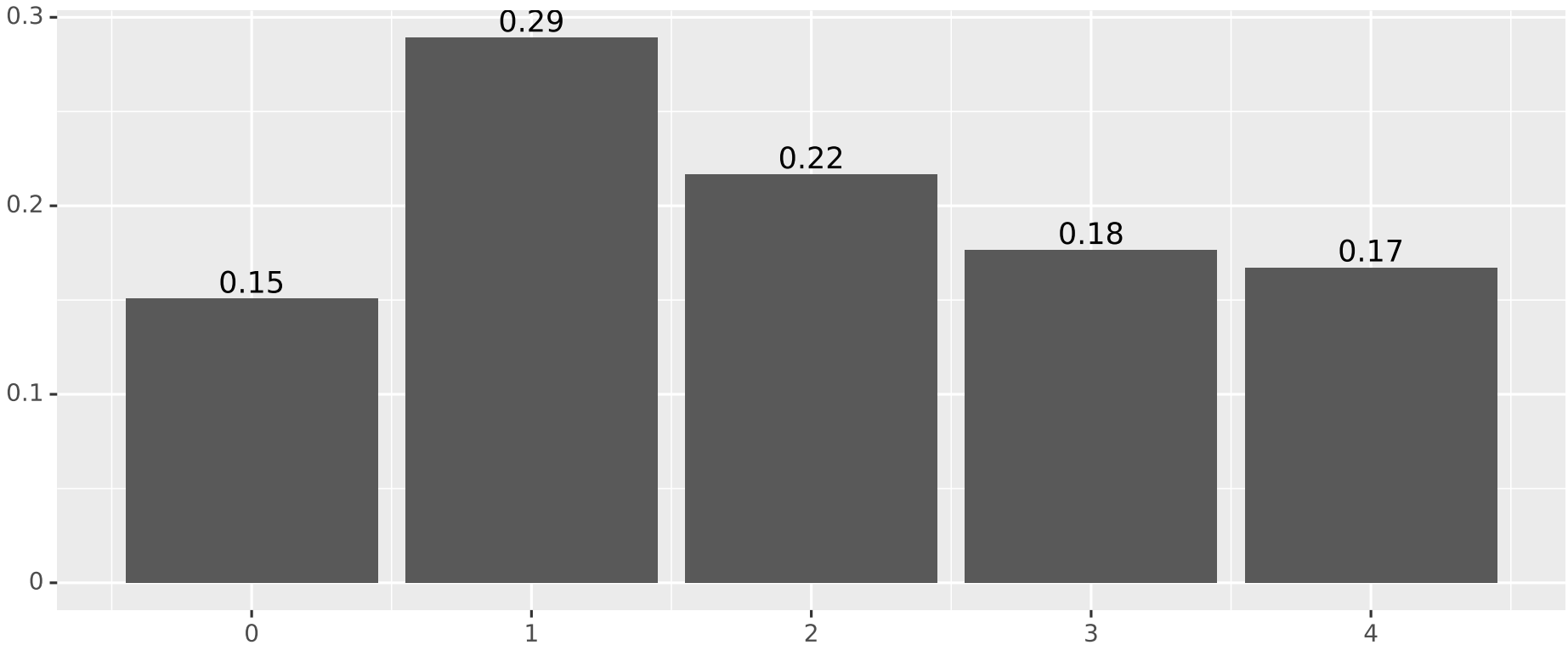
Divide into distinct groups using k-means clustering

Figure 15:



Customer segment group by percent

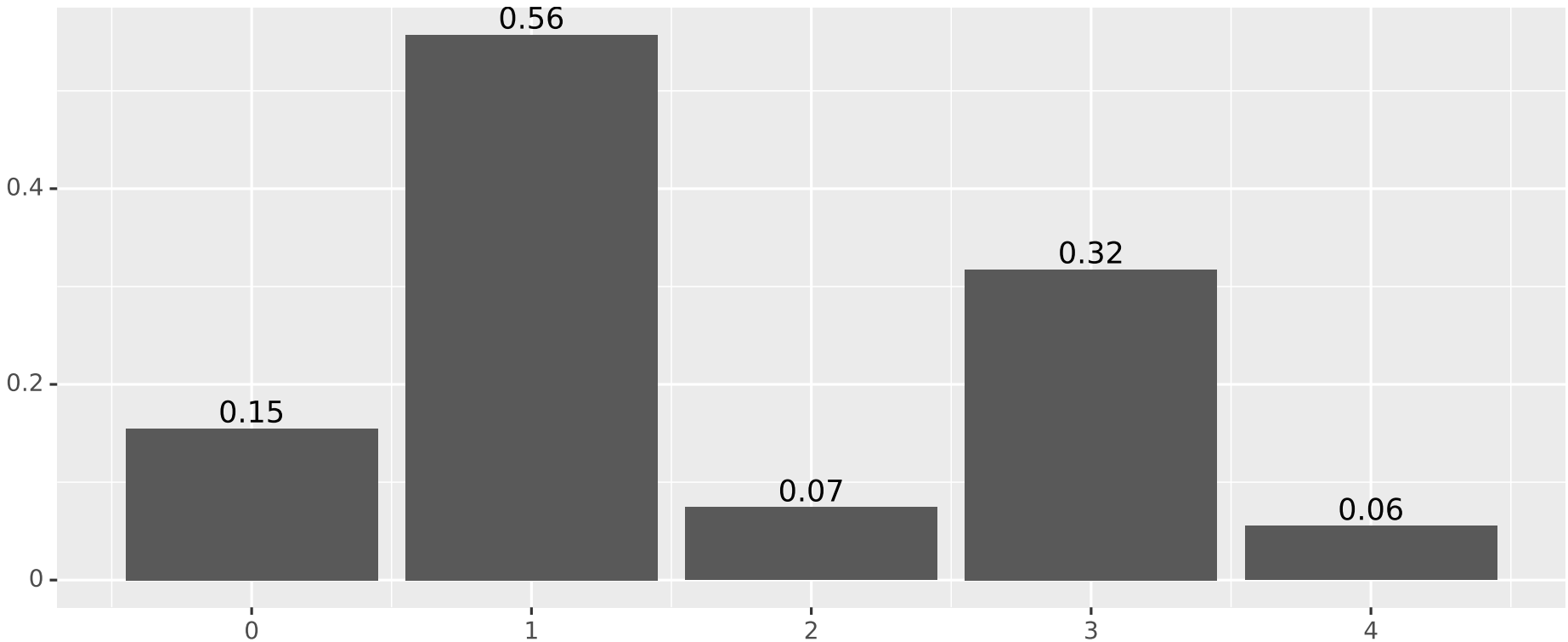
Figure 16:



Customer churn by segment group

2 Segments are most likely to churn

Figure 17:



Customer segmentation analysis

- Key point: month-to-month customers are most likely to churn
- 4 distinct groups
- 2 groups with month-to-month most likely to churn
  - 1 group has fiber optic
  - 1 group has no internet
- 2 groups without month-to-month less likely to churn
  - 1 group has fiber optic
  - 1 group has no internet

Thank you