Lecture 3 - Exploratory Data Analysis

Adnan Ferdous Ashrafi

Stamford University Bangladesh

Table of Contents

Exploratory Data Analysis

Tukey defined data analysis in 1961 as:

"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." [1]

A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

HYPOTHESIS TESTING VERSUS EXPLORATORY DATA ANALYSIS

Definition

When approaching a data mining problem, a data mining analyst may already have some a priori hypotheses that he or she would like to test regarding the relationships between the variables.

Priori Hypothesis Testing

A myriad of statistical hypothesis testing procedures are available for testing the following hypotheses:

- The Z-test for the population mean
- The t-test for the population mean
- The Z-test for the population proportion
- The Z-test for the difference in means for two populations
- The t-test for the difference in means for two populations
- The t-test for paired samples
- The Z-test for the difference in population proportions
- The X^2 goodness-of-fit test for multinomial populations
- The X^2-test for independence among categorical variables
- The analysis of variance F-test
- The t-test for the slope of the regression line

Apriori Hypothesis Testing/EDA

However, analysts do not always have a priori notions of the expected relationships among the variables. Especially when confronted with large unknown databases, analysts often prefer to use exploratory data analysis (EDA) or graphical data analysis. EDA allows the analyst to:

- Delve into the data set
- Examine the interrelationships among the attributes
- Identify interesting subsets of the observations
- Develop an initial idea of possible associations between the attributes and the target variable, if any

GETTING TO KNOW THE DATA SET

In Lecture 3 we use exploratory methods to delve into the churn data set [2] from the UCI Repository of Machine Learning Databases at the University of California, Irvine.

Churn, also called attrition, is a term used to indicate a customer leaving the service of one company in favor of another company. To begin, it is often best simply to take a look at the field values for some of the records.

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls	churn
0	KS	128	415	382-4657	no	yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False
1	OH	107	415	371-7191	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False
2	NJ	137	415	358-1921	no	no	0	243.4	114	41.38	121.2	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False
3	OH	84	408	375-9999	yes	no	0	299.4	71	50.90	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False
4	OK	75	415	330-6626	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False
5	AL	118	510	391-8027	yes	no	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.70	0	False
6	MA	121	510	355-9993	no	yes	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03	3	False
7	MO	147	415	329-9001	yes	no	0	157.0	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	0	False
8	LA	117	408	335-4719	no	no	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35	1	False
9	WV	141	415	330-8173	yes	yes	37	258.6	84	43.96	222.0	111	18.87	326.4	97	14.69	11.2	5	3.02	0	False

Figure 1: Field values of the first 10 records in the churn data set.

Variables

The data set contains 20 variables worth of information about 3333 customers, along with an indication of whether or not that customer churned (left the company). The variables are as follows:

- State: categorical, for the 50 states and the District of Columbia
- Account length: integer-valued, how long account has been active
- Area code: categorical
- Phone number: essentially a surrogate for customer ID
- International Plan: dichotomous categorical, yes or no
- Voicemail Plan: dichotomous categorical, yes or no
- Number of voice mail messages: integer-valued
- Total day minutes: continuous, minutes customer used service during the day
- Total day calls: integer-valued
- Total day charge: continuous, perhaps based on foregoing two variables
- Total evening minutes: continuous, minutes customer used service during the evening
- Total evening calls: integer-valued
- Total evening charge: continuous, perhaps based on foregoing two variables
- Total night minutes: continuous, minutes customer used service during the night
- Total night calls: integer-valued
- Total night charge: continuous, perhaps based on foregoing two variables
- Total international minutes: continuous, minutes customer used service to make international calls
- Total international calls: integer-valued
- Total international charge: continuous, perhaps based on foregoing two variables
- Number of calls to customer service: integer-valued

Co-related Variables

One should take care to avoid feeding correlated variables to one's data mining and statistical models. At best, using correlated variables will overemphasize one data component; at worst, using correlated variables will cause the model to become unstable and deliver unreliable results.

The data set contains three variables: total_day_minutes, total_day_calls, and total_day_charge. The data description indicates that the charge variable may be a function of minutes and calls, with the result that the variables would be correlated. We investigate using the correlation matrix plot shown in Figure 2.

Co-related Variables - Correlation Matrix

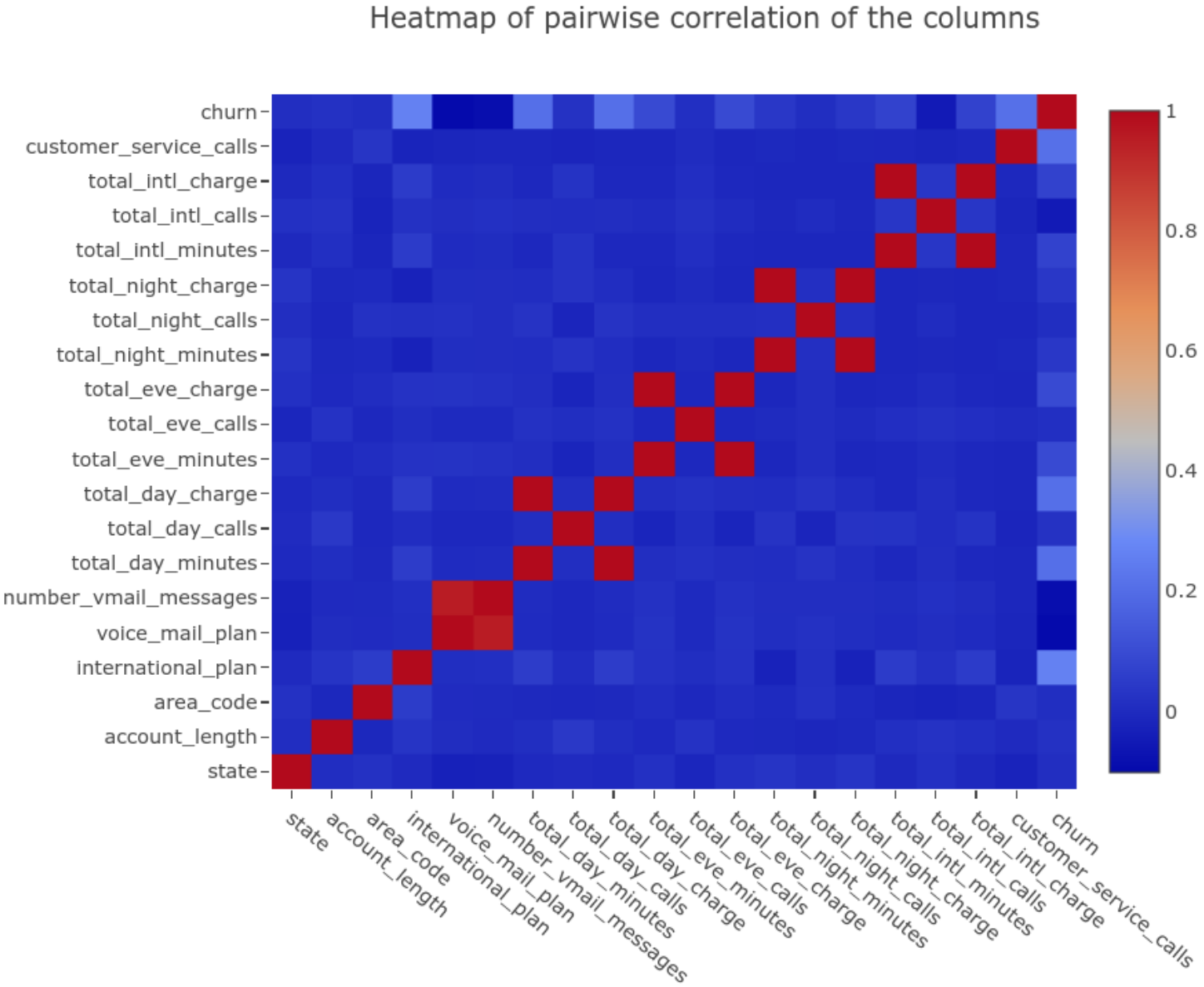


Figure 2: Heatmap of pairwise correlation of the columns.

Inference- Inference

1. There does not seem to be any relationship between day minutes and day calls or between day calls and day charge.
2. On the other hand, there is a perfect linear relationship between total day minutes and total day charge, indicating that day charge is a simple linear function of day minutes only.
3. Since day charge is correlated perfectly with day minutes, we should eliminate one of the two variables.
4. Investigation of the evening, night, and international components reflected similar findings, and we thus also eliminate evening charge, night charge, and international charge.
5. Dimensionality of the solution space is reduced, so that certain data mining algorithms may more efficiently find the globally optimal solution.

We have therefore reduced the number of predictors from 20 to 16 by eliminating redundant variables.

EXPLORING CATEGORICAL VARIABLES

One of the primary reasons for performing exploratory data analysis is to investigate the variables, look at histograms of the numeric variables, examine the distributions of the categorical variables, and explore the relationships among sets of variables.

EXPLORING CATEGORICAL VARIABLES- Histograms

For example, Figure 3 shows that we have clearly more samples for customers without churn than for customers with churn. So we have a class imbalance for the target variable which could lead to predictive models which are biased towards the majority (i.e. no churn). In order to deal with this issue we will investigate into the use of oversampling when building the models.

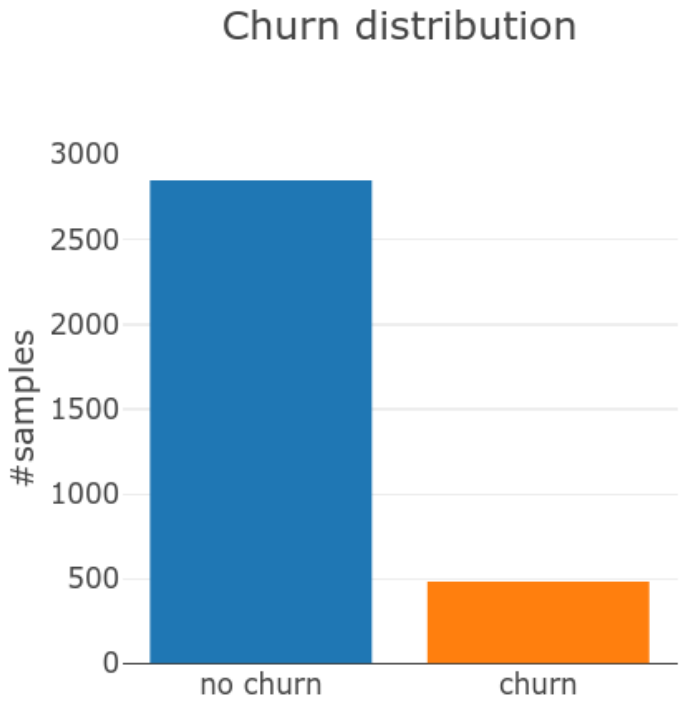


Figure 3: Bar chart of distribution of the our target variable.

EXPLORING CATEGORICAL VARIABLES- Histograms

For example, Figure 4 shows a comparison of the proportion of churners (orange) and non-churners (blue) among customers who either had selected the International Plan (yes, 346 of customers) or had not selected it (no, 2664 of customers). The graphic appears to indicate that a greater proportion of International Plan holders are churning, but it is difficult to be sure.

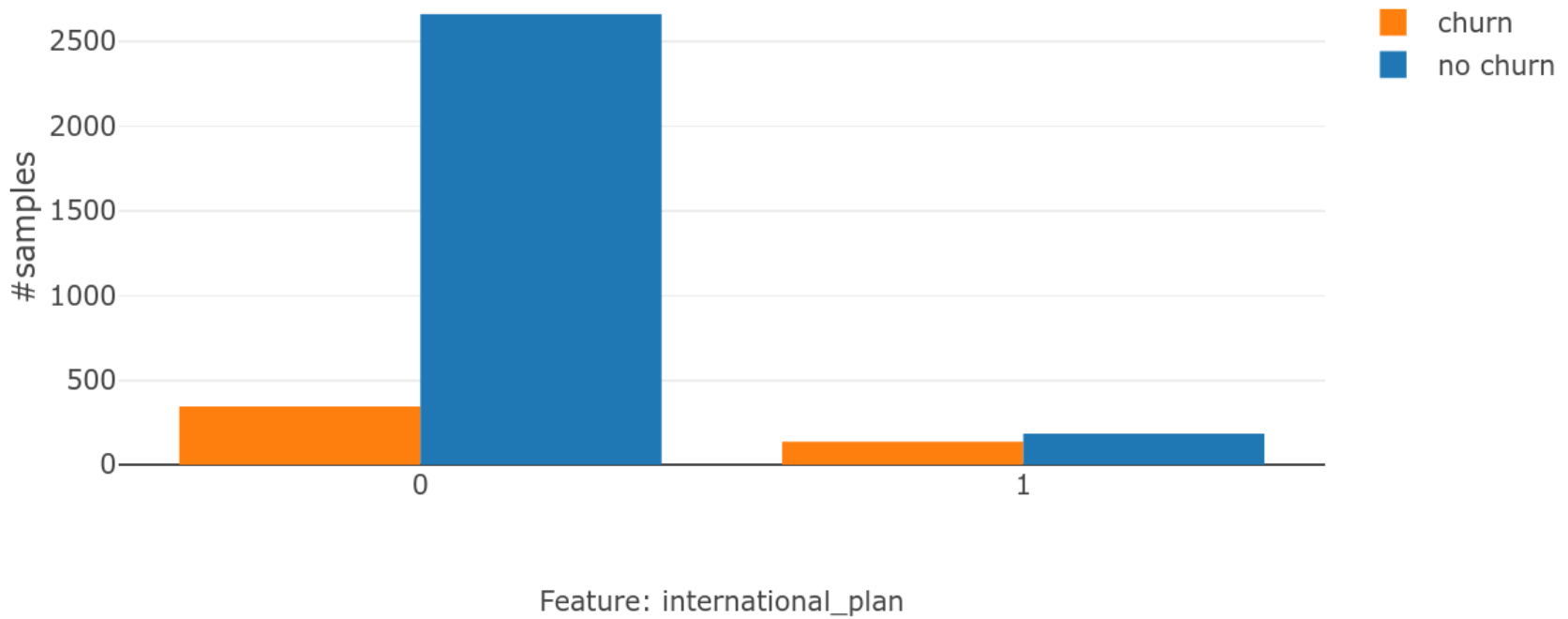


Figure 4: Comparison bar chart of churn proportions by Intl. Plan participation.

EXPLORING CATEGORICAL VARIABLES- Histograms

Note that 137/(137 + 186) = 42.4% of the International Plan holders churned, compared with only 346/(346 + 2664) = 11.5% of those without the International Plan. Customers selecting the International Plan are more than three times as likely to leave the company's service than those without the plan.

This EDA on the International Plan has indicated that:

1. Perhaps we should investigate what it is about the International Plan that is inducing customers to leave!
2. We should expect that whatever data mining algorithms we use to predict churn, the model will probably include whether or not the customer selected the International Plan.

EXPLORING CATEGORICAL VARIABLES- Histograms

Let us now turn to the VoiceMail Plan. Figure 5 shows in a bar graph that those who do not have the VoiceMail Plan are more likely to churn than those who do have the plan.

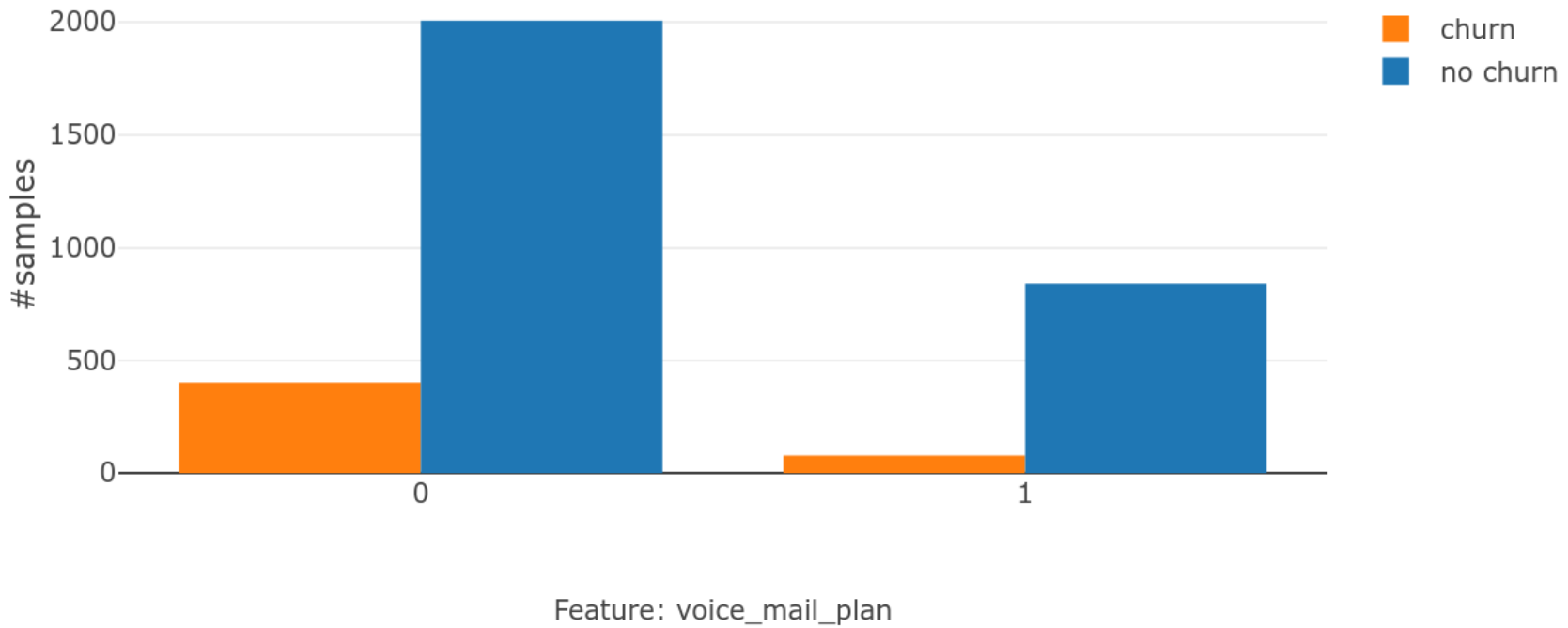


Figure 5: Those without the VoiceMail Plan are more likely to churn.

EXPLORING CATEGORICAL VARIABLES- Histograms

First of all, $842 + 80 = 922$ customers have the VoiceMail Plan, while $2008 + 403 = 2411$ do not. We then find that $403/2411 = 16.7\%$ of those without the VoiceMail Plan are churners, compared to $80/922 = 8.7\%$of customers who do have the VoiceMail Plan. Thus, customers without the VoiceMail Plan are nearly twice as likely to churn as customers with the plan.

This EDA on the International Plan has indicated that:

1. Perhaps we should enhance the VoiceMail Plan further or make it easier for customers to join it, as an instrument for increasing customer loyalty.
2. We should expect that whatever data mining algorithms we use to predict churn, the model will probably include whether or not the customer selected the VoiceMail Plan. Our confidence in this expectation is perhaps not quite as the high as that for the International Plan.

EXPLORING CATEGORICAL VARIABLES- Histograms

For example, in Figure 6 we can see that some states have less proportion of customer with churn like AK, HI, IA and some have a higher proportion such as WA, MD and TX. This shows that we should incorporate the state into our further analysis, because it could be help to predict if a customer is going to churn.

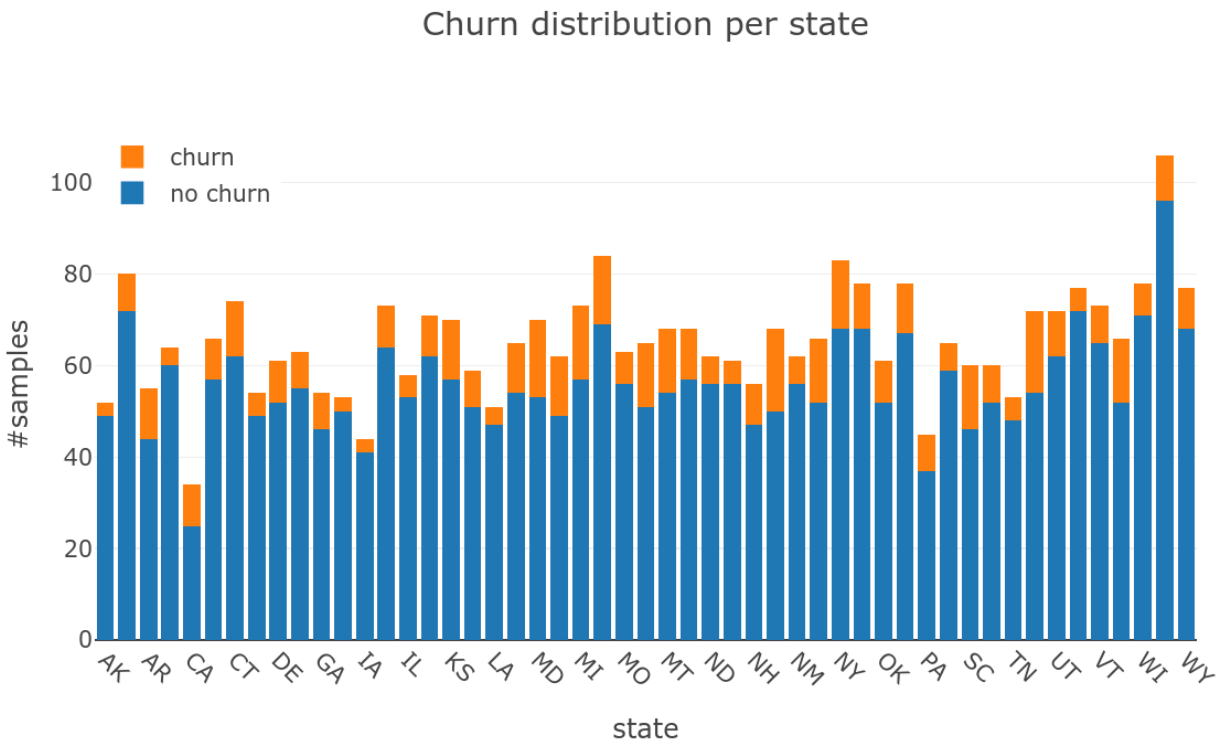
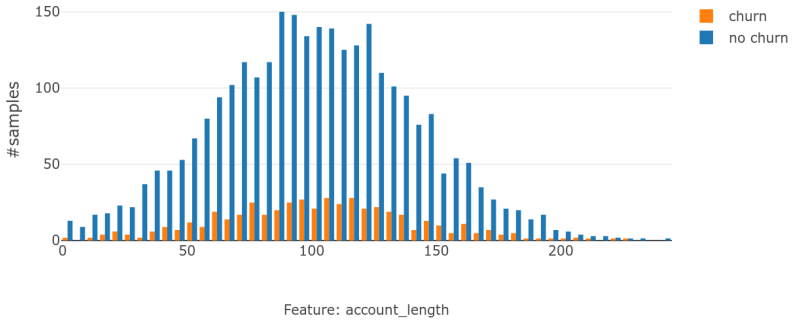
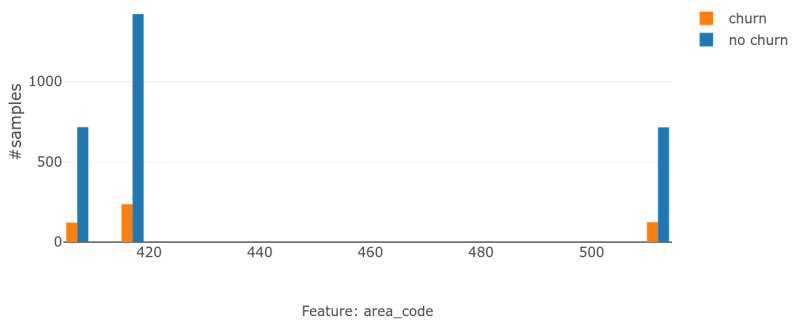


Figure 6: How much the state influences our target?

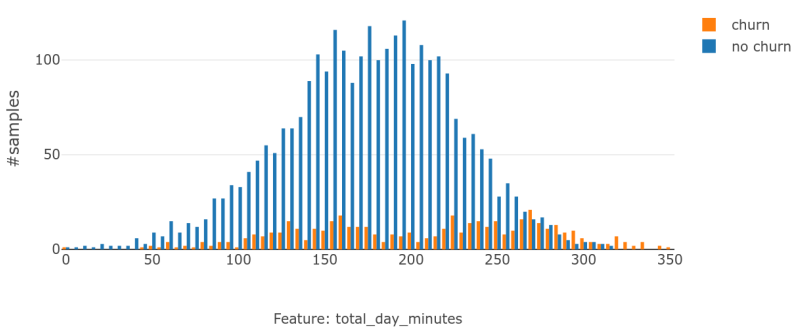
Can you find a pattern?



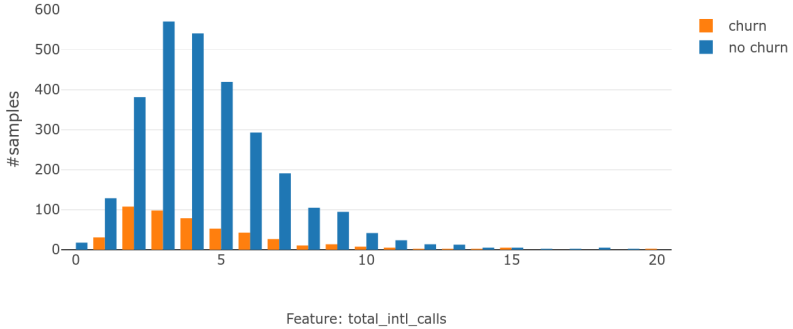
(a) Comparison bar chart of churn proportions by Account Length.



(b) Comparison bar chart of churn proportions by Area Code.



(c) Comparison bar chart of churn proportions by Total Day Minutes.



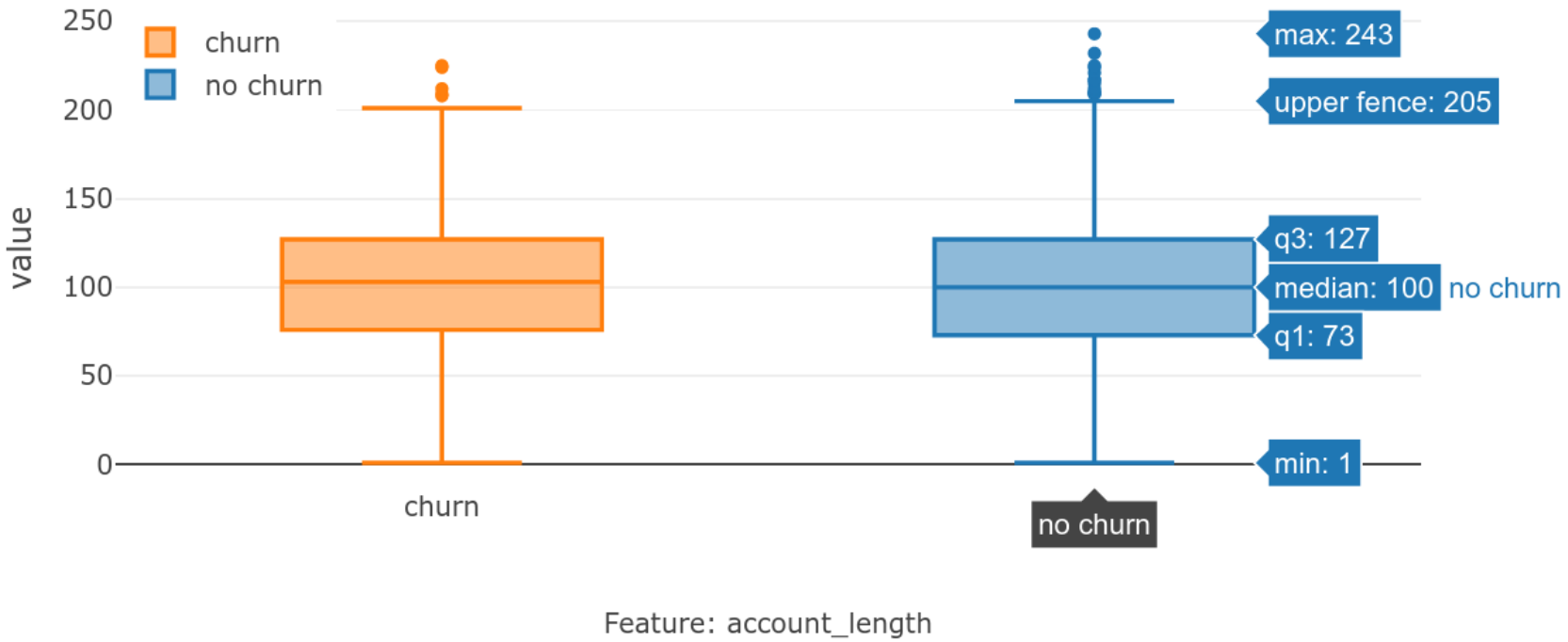
(d) Comparison bar chart of churn proportions by No. of Intl. Calls.

Figure 7: Four distribution graphs. Can you find some inference from here?

EXPLORING CATEGORICAL VARIABLES- Distribution Box Plots

Next, we take a look at the box plots for each feature. A box plot visualizes the following statistics:

- median
- the first quartile (Q1) and the third quartile (Q3) building the interquartile range (IQR)
- the lower fence (Q1 - 1.5 IQR) and the upper fence (Q3 + 1.5 IQR)
- the maximum and the minimum value



EXPLORING CATEGORICAL VARIABLES- Distribution Box Plots

Looking at the account length , the box plot shows that both churn and no-churn customers have a similar amount of account length, and both type of users have similar median of account length.

This EDA on the Account Length has indicated that:

1. Perhaps there are no differences in case of account length for both churn and non-churn customers!
2. Both has almost identical median, IQR, min and max values

EXPLORING CATEGORICAL VARIABLES- Distribution Box Plots

When we look at the box plot in Figure 9 for the number of voice mail messages ("number vmail messages"), we can see that we have some outliers for the customers with churn, but most of them have send zero voice mail messages. The customers which did not churn instead tend to do more voice mail messages.

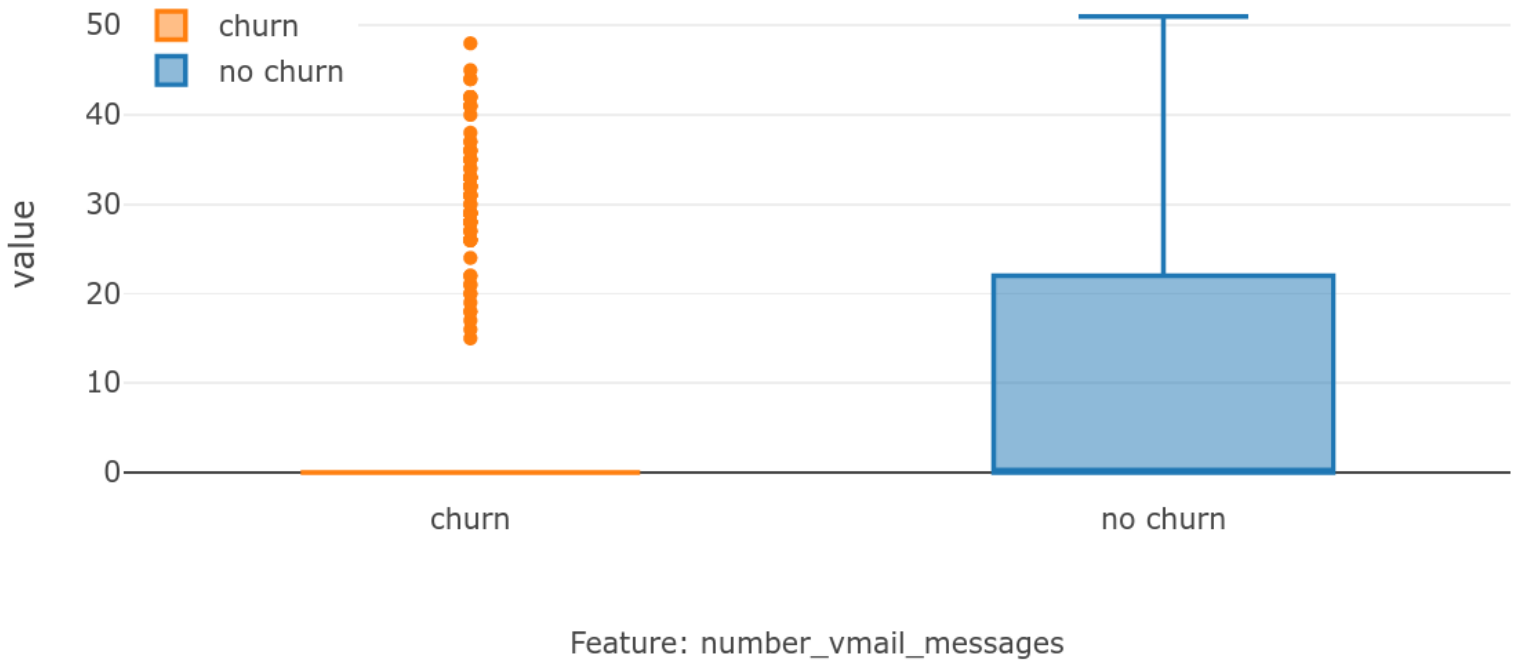
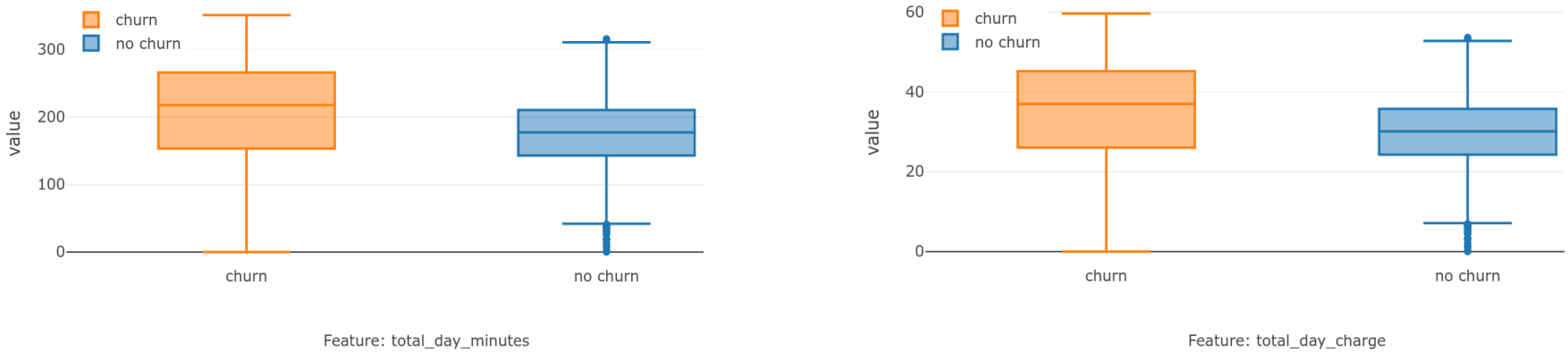


Figure 9: Box plot for the number of voicemail messages.

EXPLORING CATEGORICAL VARIABLES- Distribution Box Plots

In Figure 10 Similar to our findings in the histograms, we can see also in the box plot that the median of the total day minutes and the total day charge for churn clients is higher than the one of no-churn clients.



(a) Box plot for the total day minutes.

(b) Box plot for the total day charge.

Figure 10: Box plots show similar idea to their respective histograms.

EXPLORING CATEGORICAL VARIABLES- Distribution Box Plots

Looking at the Figure 11 of total international calls, the box plot shows that both churn and no-churn customers are doing a similar amount of international calls, but the churn-customers tend to do longer calls as the median of churn customers for the total international minutes is higher than for the no-churn customers.

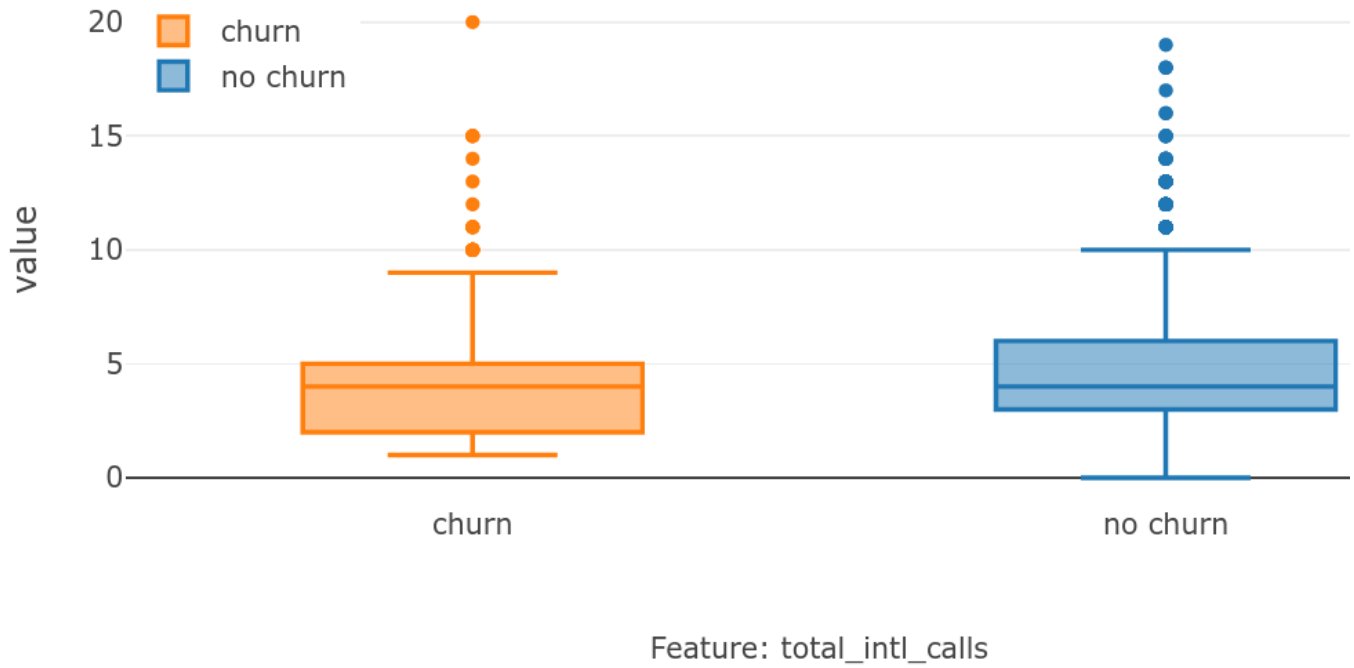


Figure 11: Box plot for the total international minutes called.

EXPLORING CATEGORICAL VARIABLES- Distribution Box Plots

Finally, the plot for the number of customer service calls shows that clients with churn have a higher median and a higher variance for the customer service calls.

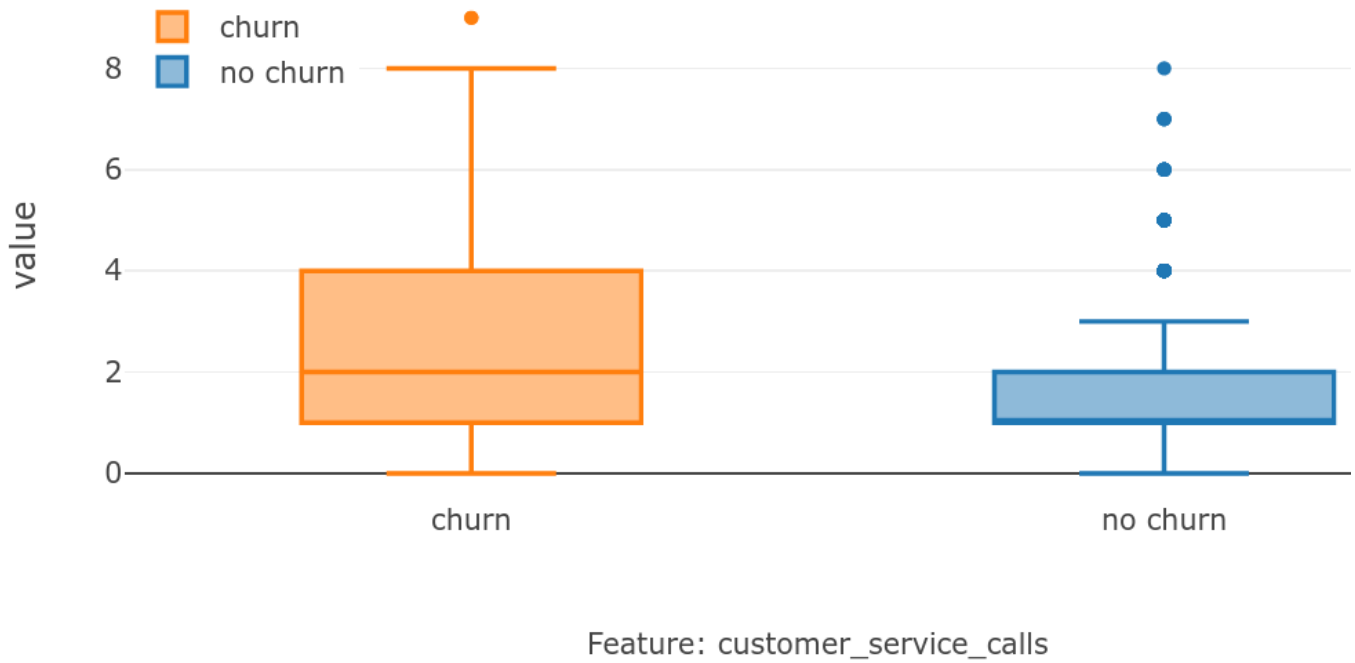
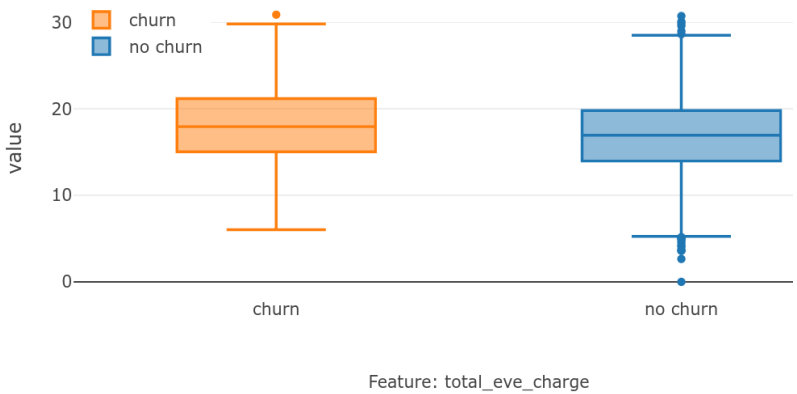
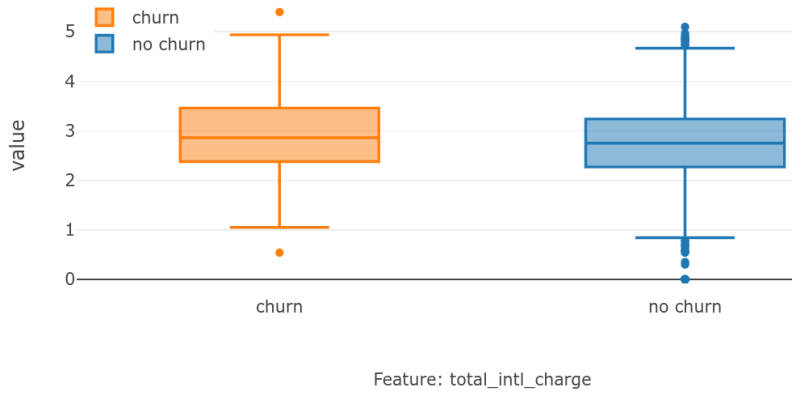


Figure 12: Box plot for the total service calls made by customers.

Can you find a pattern?



(a) Box plot for the total evening charge of customers.



(b) Box plot for the total international charge of customers.

Figure 13: Two box plot graphs. Can you find some inference from here?

USING EDA TO UNCOVER ANOMALOUS FIELDS

Exploratory data analysis will sometimes uncover strange or anomalous records or fields which the earlier data cleaning phase may have missed. Consider, for example, the area code field in the present data set. Although the area codes contain numerals, they can also be used as categorical variables, since they can classify customers according to geographical location.

USING EDA TO UNCOVER ANOMALOUS FIELDS- Histograms

We are intrigued by the fact that the area code field contains only three different values for all the records—408, 415, and 510—all three of which are in California, as shown by Figure 14.

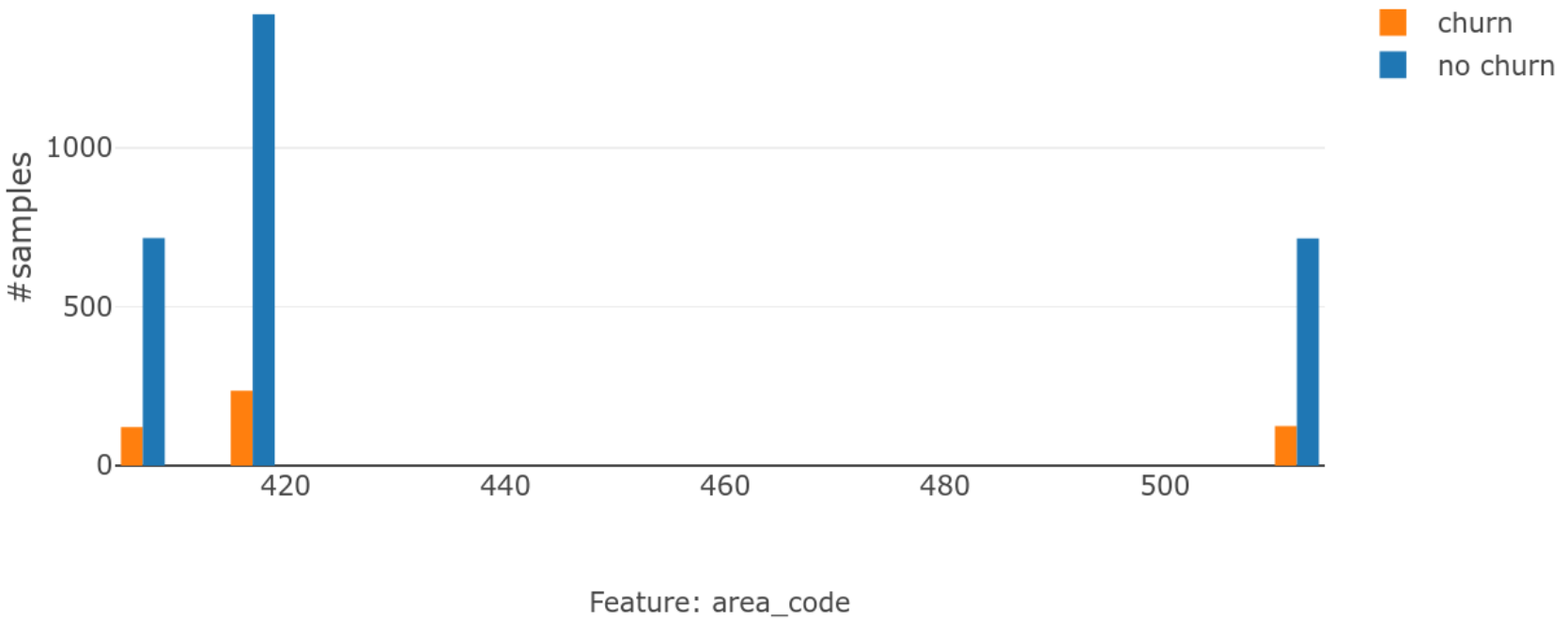


Figure 14: Only three area codes for all records. Why do we need the others?

USING EDA TO UNCOVER ANOMALOUS FIELDS- Histograms

Now, this would not be anomalous if the records indicated that the customers all lived in California. However, as shown in the cross-tabulation in Figure 15, the three area codes seem to be distributed more or less evenly across all the states and the District of Columbia. It is possible that domain experts might be able to explain this type of behavior, but it is also possible that the field just contains bad data.

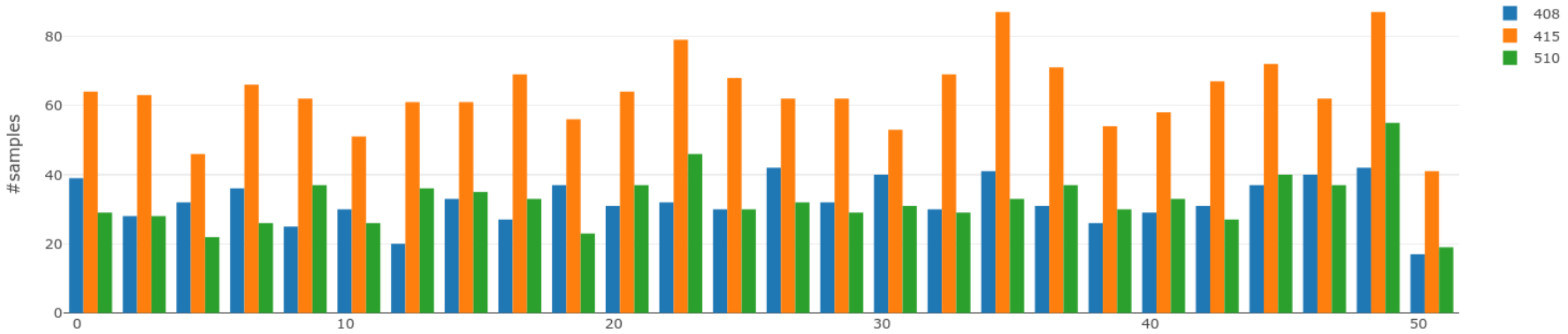


Figure 15: Anomaly: three area codes distributed across all 50 states.

USING EDA TO UNCOVER ANOMALOUS FIELDS- Inference

We should therefore be wary of this *area code* field, perhaps going so far as not to include it as input to the data mining models in the next phase. On the other hand, it may be the *state* field that is in error. Either way, further communication with someone familiar with the data history, or a domain expert, is called for before inclusion of these variables in the data mining models.

EXPLORING NUMERICAL VARIABLES

Next, we turn to an exploration of the numerical predictive variables. We begin with numerical summary measures, including minimum and maximum; measures of center, such as mean, median, and mode; and measures of variability, such as standard deviation. Figure 16 shows these summary measures for some of our numerical variables. We see, for example, that the minimum account length is one month, the maximum is 243 months, and the mean and median are about the same, at around 101 months, which is an indication of symmetry. Notice that several variables show this evidence of symmetry, including all the minutes, charge, and call fields.

EXPLORING NUMERICAL VARIABLES- Statistics summary

	state	account_length	area_code	international_plan	voice_mail_plan	number_vmail_messages
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	26.059406	101.064806	437.182418	0.096910	0.276628	8.099010
std	14.824911	39.822106	42.371290	0.295879	0.447398	13.688365
min	0.000000	1.000000	408.000000	0.000000	0.000000	0.000000
25%	14.000000	74.000000	408.000000	0.000000	0.000000	0.000000
50%	26.000000	101.000000	415.000000	0.000000	0.000000	0.000000
75%	39.000000	127.000000	510.000000	0.000000	1.000000	20.000000
max	50.000000	243.000000	510.000000	1.000000	1.000000	51.000000

Figure 16: Summary statistics for several numerical variables.

EXPLORING NUMERICAL VARIABLES- Histograms

We turn next to graphical analysis of our numerical variables. Figure 17 is a histogram of customer service calls, with churn overlay. Figure 17 hints that the proportion of churn may be greater for higher numbers of customer service calls, but it is difficult to discern this result unequivocally.

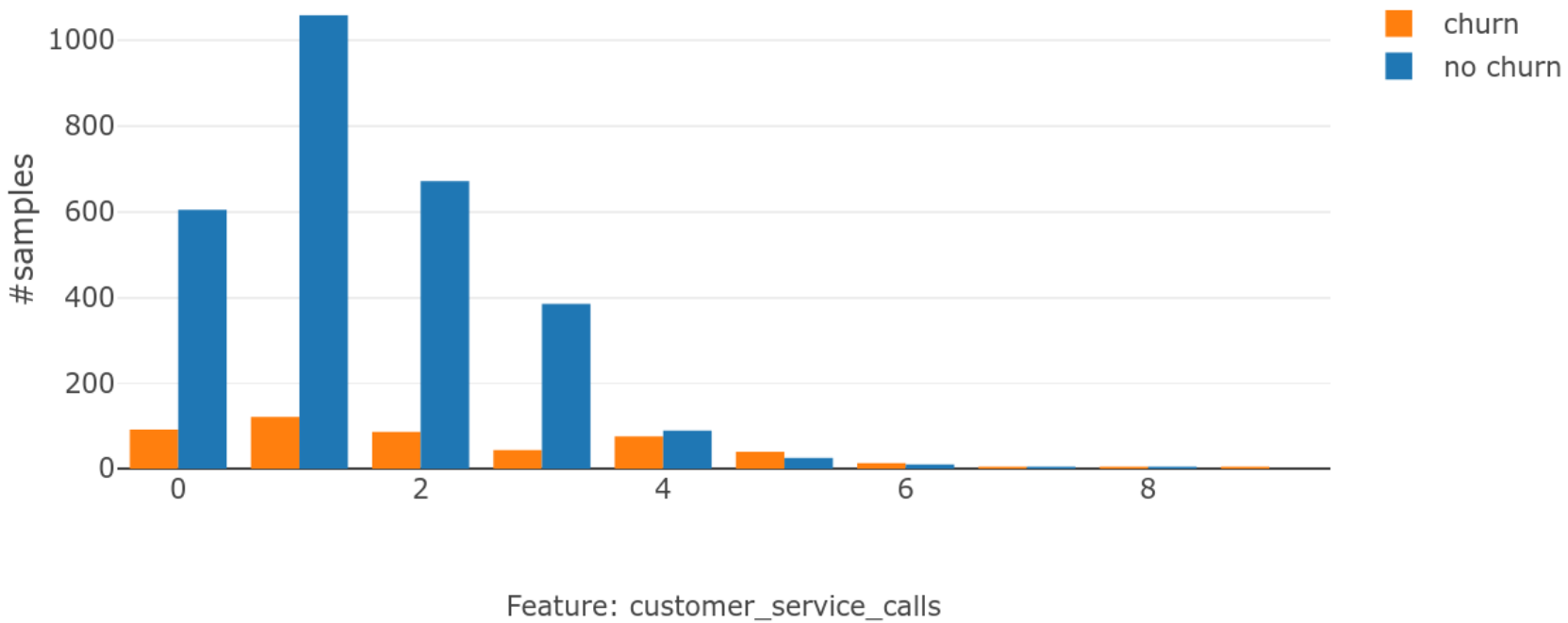


Figure 17: Histogram of customer service calls, with churn overlay.

EXPLORING NUMERICAL VARIABLES- Inference

This EDA on the customer service calls has indicated that:

1. We should track carefully the number of customer service calls made by each customer. By the third call, specialized incentives should be offered to retain customer loyalty.
2. We should expect that whatever data mining algorithms we use to predict churn, the model will probably include the number of customer service calls made by the customer.

EXPLORING NUMERICAL VARIABLES- Histograms

Examining Figure 3.19, we see that the normalized histogram of day minutes indicates that very high day users tend to churn at a higher rate. Therefore:

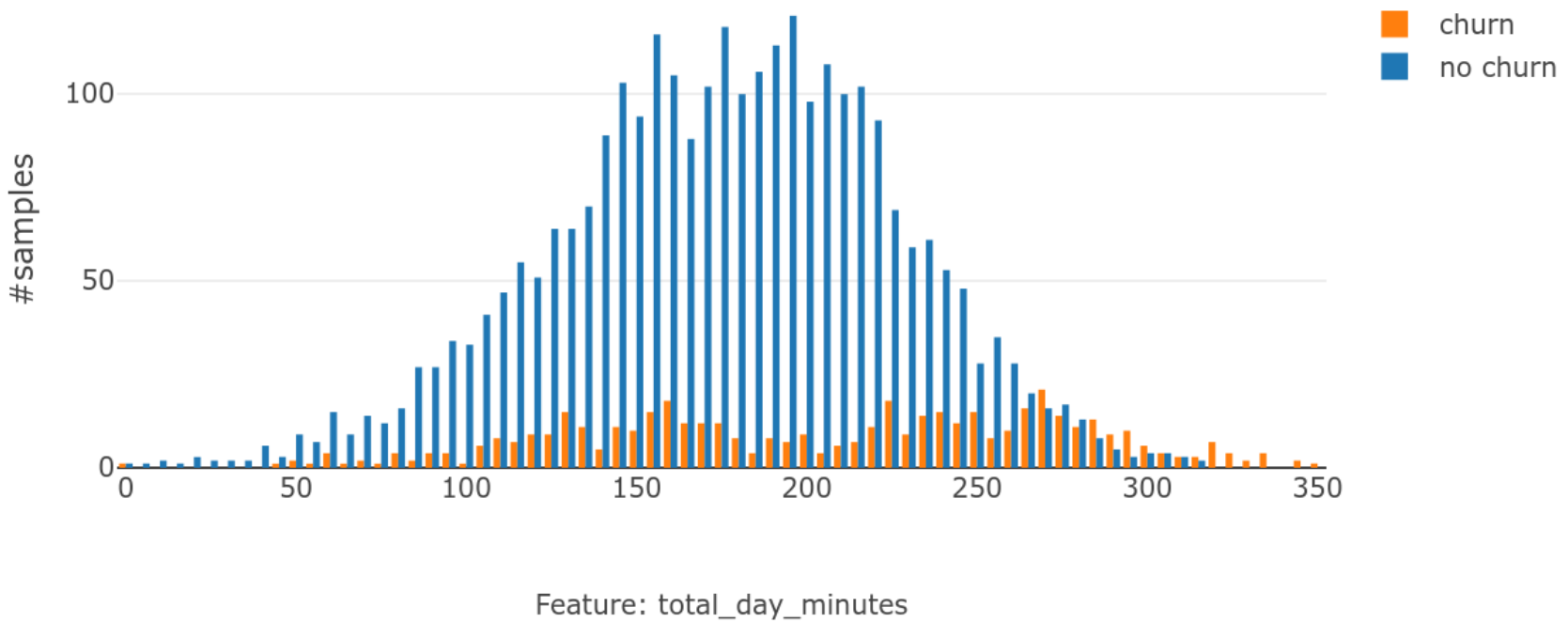


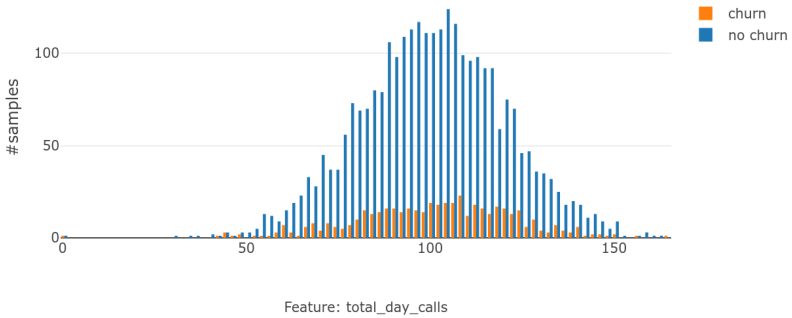
Figure 18: Customers with high day minutes tend to churn at a higher rate.

EXPLORING NUMERICAL VARIABLES- Inference

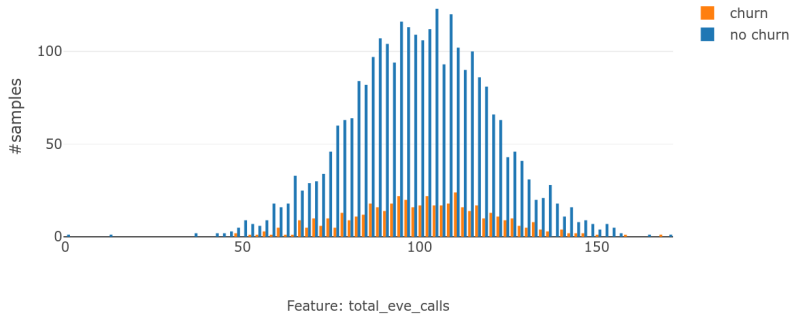
This EDA on the customer service calls has indicated that:

1. We should carefully track the number of day minutes used by each customer. As the number of day minutes passes 200, we should consider special incentives.
2. We should investigate why heavy day users are tempted to leave.
3. We should expect that our eventual data mining model will include day minutes as a predictor of churn.

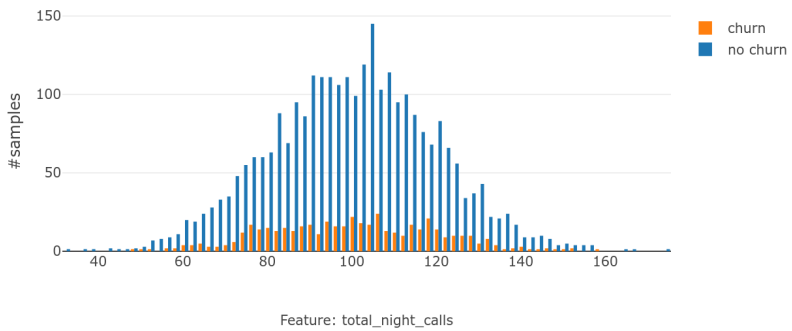
EXPLORING NUMERICAL VARIABLES- Histograms



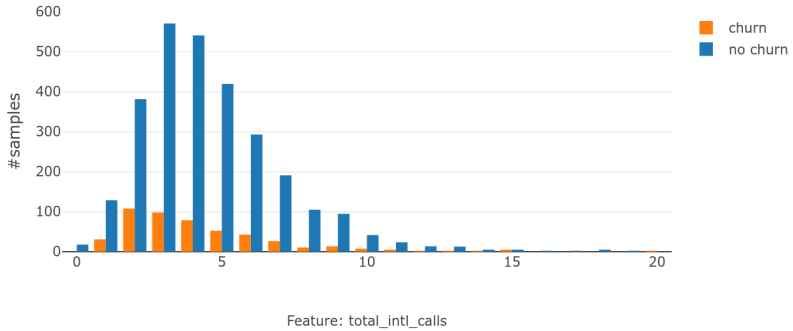
(a) Comparison bar chart of churn proportions by Day Calls.



(b) Comparison bar chart of churn proportions by Evening Calls.



(c) Comparison bar chart of churn proportions by Night Calls.



(d) Comparison bar chart of churn proportions by International Calls.

Figure 19: No association of churn with day calls, evening calls, night calls, or international calls.

Summary After Numerical and Categorical Analysis

Variable	Disposition
State	Anomalous. Omitted from model.
Account length	No obvious relation with churn, but retained.
Area code	Anomalous. Omitted from model.
Phone number	Surrogate for ID. Omitted from model.
International Plan	Predictive of churn. Retained.
VoiceMail Plan	Predictive of churn. Retained.
Number of voice mail messages	No obvious relation with churn, but retained.
Total day minutes	Predictive of churn. Retained.
Total day calls	No obvious relation with churn, but retained.
Total day charge	Function of *minutes*. Omitted from model.
Total evening minutes	May be predictive of churn. Retained.
Total evening calls	No obvious relation with churn, but retained.
Total evening charge	Function of *minutes*. Omitted from model.
Total night minutes	No obvious relation with churn, but retained.
Total night calls	No obvious relation with churn, but retained.
Total night charge	Function of *minutes*. Omitted from model.
Total international minutes	No obvious relation with churn, but retained.
Total international calls	No obvious relation with churn, but retained.
Total international charge	Function of *minutes*. Omitted from model.
Customer service calls	Predictive of churn. Retained.

Figure 20: Summary of Exploratory Findings Thus Far.

EXPLORING MULTIVARIATE RELATIONSHIPS - Scatter-plots

We turn next to an examination of possible multivariate associations of numerical variables with churn, using two- and three-dimensional scatter plots.

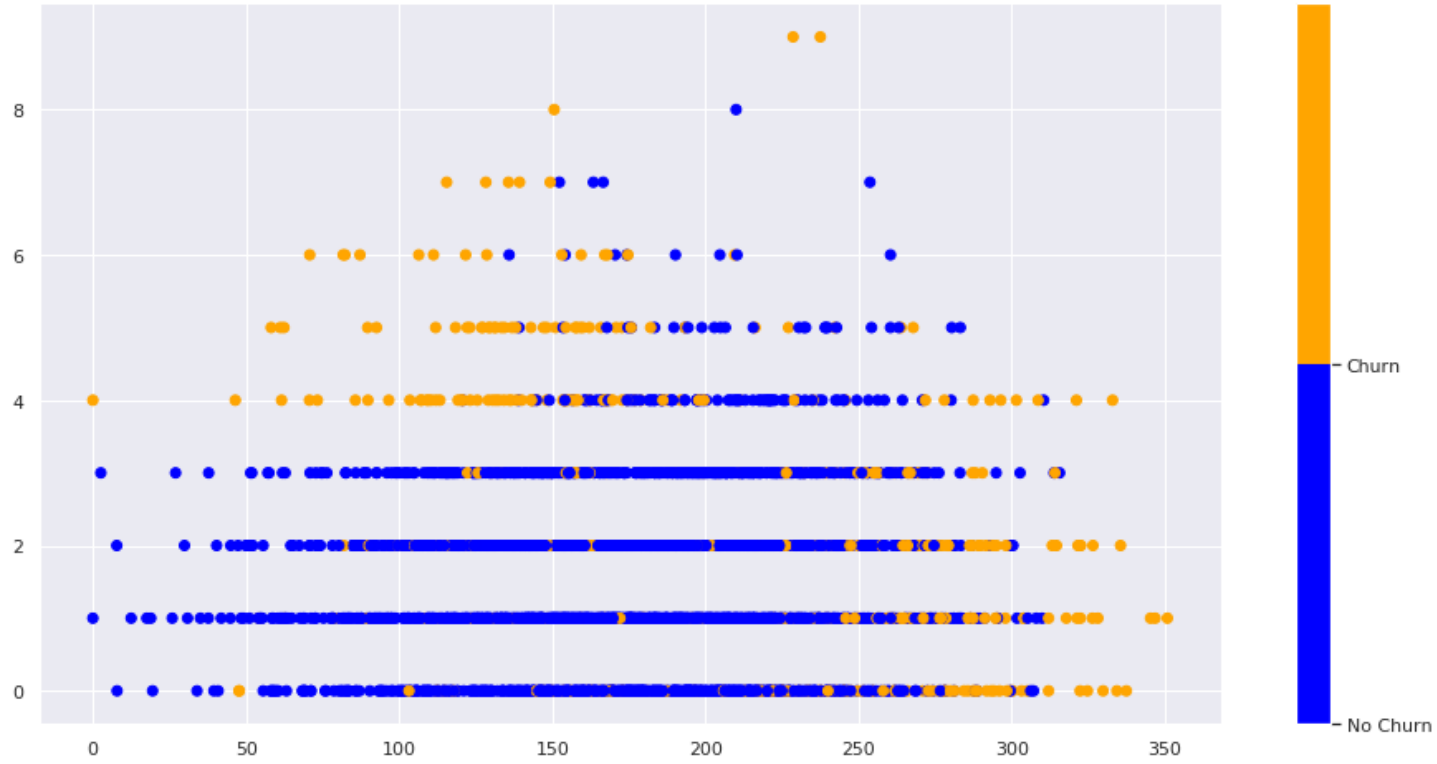


Figure 21: Scatter plot of customer service calls versus day minutes with churn overlay.

EXPLORING MULTIVARIATE RELATIONSHIPS - Scatter-plots

Figure 21 is a scatter plot of customer service calls versus account length. This shows that of these customers with high numbers of customer service calls, those who also have high day minutes are somewhat “protected” from this high churn rate. The customers in the upper right of the scatter plot exhibit a lower churn rate than that of those in the upper left.

EXPLORING MULTIVARIATE RELATIONSHIPS - Scatter-plots

Sometimes, three-dimensional scatter plots can be helpful as well.

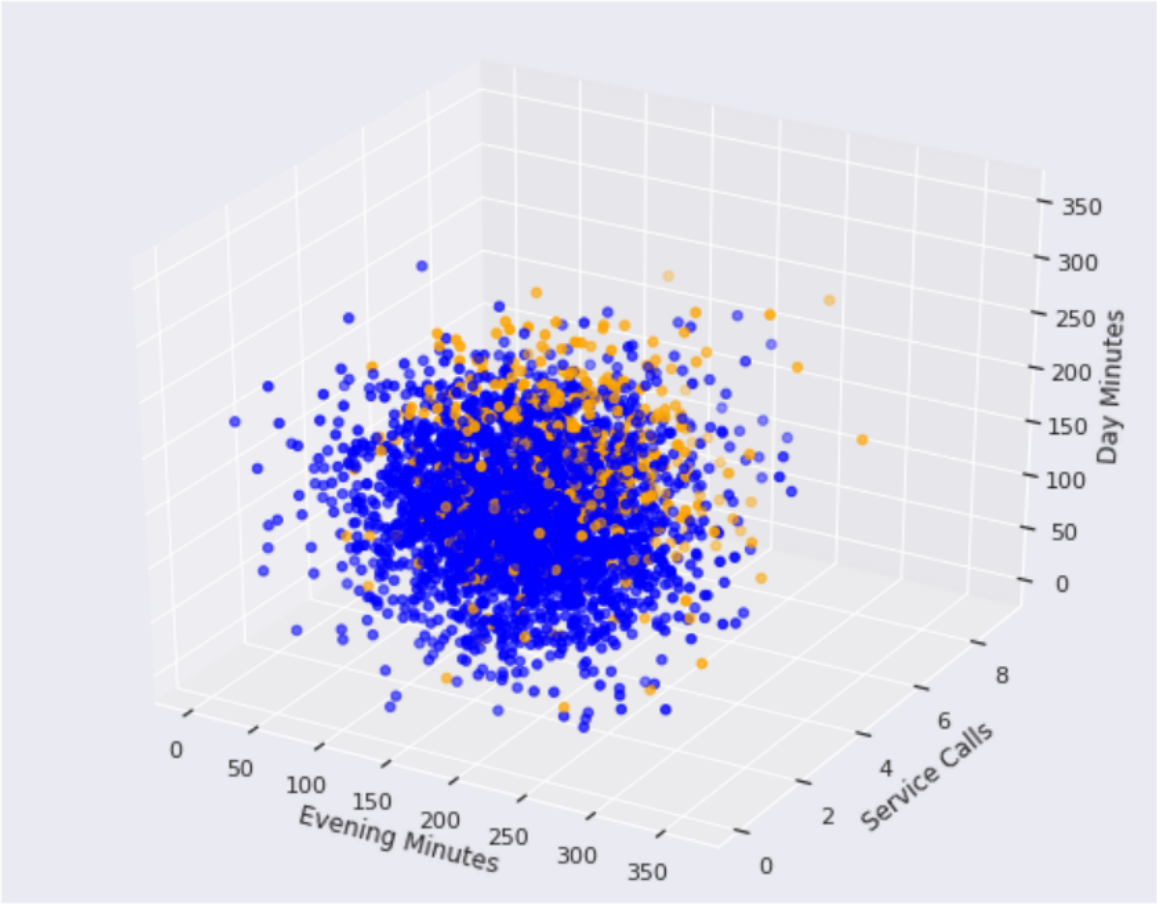


Figure 22: Three-dimensional scatter plot of day minutes versus evening minutes versus customer service calls, with a churn overlay.

EXPLORING MULTIVARIATE RELATIONSHIPS - Scatter-plots

Figure 22 is an example of a plot of day minutes versus evening minutes versus customer service calls, with a churn overlay. The scroll buttons on the sides rotate the display so that the points may be examined in a three-dimensional environment.

SELECTING INTERESTING SUBSETS OF THE DATA FOR FURTHER INVESTIGATION

We may use scatter plots (or histograms) to identify interesting subsets of the data, in order to study these subsets more closely. In Figure 3.25 we see that customers with high day minutes and high evening minutes are more likely to churn.

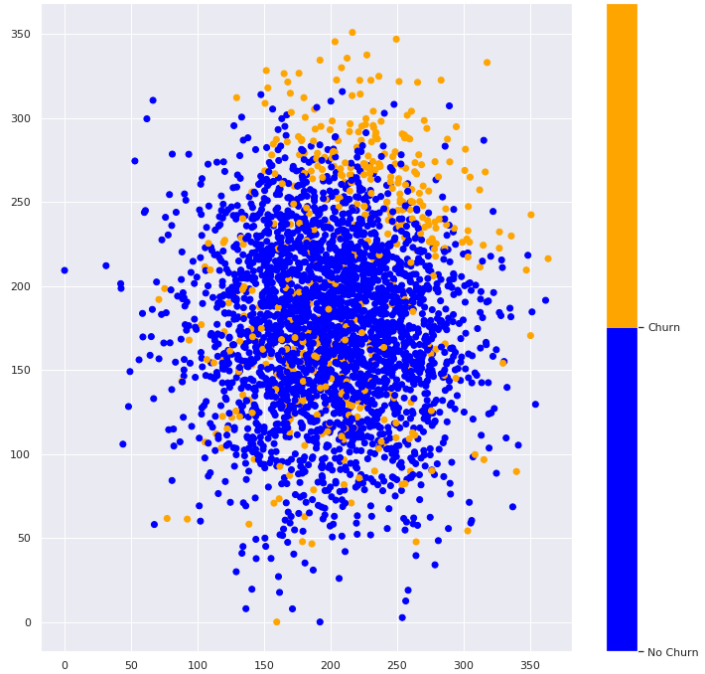


Figure 23: Selecting an interesting subset of records for further investigation.

SELECTING INTERESTING SUBSETS OF THE DATA FOR FURTHER INVESTIGATION - Inference

It turns out that over 43% of the customers who have both high day minutes and high evening minutes are churners. This is approximately three times the churn rate of the overall customer base in the data set. Therefore, it is recommended that we consider how we can develop strategies for keeping our heavy-use customers happy so that they do not leave the company's service, perhaps through discounting the higher levels of minutes used.

BINNING

Definition

Binning (also called banding) refers to the categorization of numerical or categorical variables into a manageable set of classes which are convenient for analysis.

For example, the number of day minutes could be categorized (binned) into three classes: low, medium, and high. The categorical variable state could be binned into a new variable, *region*, where California, Oregon, Washington, Alaska, and Hawaii would be put in the Pacific category, and so on. Properly speaking, binning is a data preparation activity as well as an exploratory activity.

Binning

There are various strategies for binning numerical variables. One approach is to make the classes of equal width, analogous to equal-width histograms. Another approach is to try to equalize the number of records in each class. You may consider yet another approach, which attempts to partition the data set into identifiable groups of records, which, with respect to the target variable, have behavior similar to that for other records in the same class.

Binning

For example, recall Figure 17, where we saw that customers with fewer than four calls to customer service had a lower churn rate than that of customers who had four or more calls to customer service. We may therefore decide to bin the customer service calls variable into two classes, low and high.

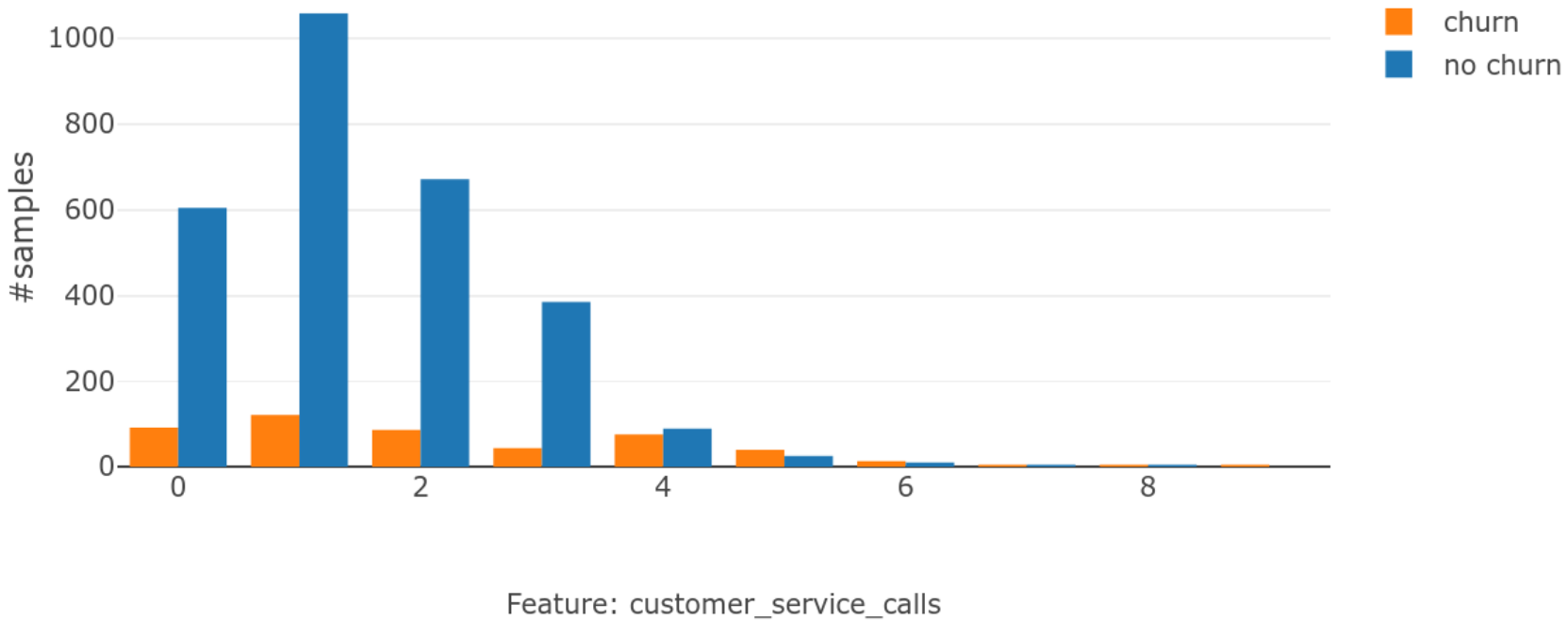


Figure 24: Churn rate for customers with low (top) and high (bottom) customer service calls.

Further Reading

- Chapter 3 of Data Mining - Practical Machine Learning Tools and Techniques, Second Edition - Ian H. Witten, Eibe Frank
- Chapter 3 of DISCOVERING KNOWLEDGE IN DATA - An Introduction to Data Mining - DANIEL T. LAROSE
- Chapter 3 of Introduction to Data Mining (Second Edition) - Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar

References

[1] J. W. Tukey, "The future of data analysis," *The Annals of Mathematical Statistics*, vol. 33, no. 1, pp. 1–67, 1962. [Online]. Available: http://www.jstor.org/stable/2237638

[2] R. Jafari-Marandi, J. Denton, A. Idris, B. K. Smith, and A. Keramati, "Optimum profit-driven churn decision making: innovative artificial neural networks in telecom industry," *Neural Computing and Applications*, vol. 32, 9 2020.

Thank you.
Any Questions?