20592 Statistics and Probability - Final Project

Telco Sector's Retention Commercial Campaign - Churn Phenomenon Analysis

Group M:
Knezevic Ivan
Li Fei
Ma Qitian
Pacchiana Luca
Ye Weiting

Agenda

- Objectives and Scope
- Exploratory Data Analysis (EDA)
- Bivariate analysis & data visualization
- Scoring Model - Logistic Regression
- Conclusions - Business Case Simulations

We would like to analyze Churn phenomenon on data coming from Telco Sector and propose the best Target for a Retention Commercial Campaign.

- Objectives:
- Identify the Target size (in terms of number of customers) to contact through the Retention Commercial Campaign
- Balance the costs for contacts and the revenues obtained by retained customers
- Fit a logistic regression model through PCA and necessary transformations
- Select a threshold of predicted probability (score)
- Maximize True Positive Cases (the number of customers who will churn and are contacted) under the constraint:

$\{25 * TPr * churn\_rate * 25 - 5 * [(TPr * churn\_rate + FPr * (1 - churn\_rate)] \} >= 0$

- Assess the quality of the estimates with parametric and non-parametric methods

- Scope:
- Cost for each single contact: 5 euros
- Expected Revenues for each retained customer: 25 euros
- Expected Retention rate obtained through the Commercial Campaign: 25%
- Customers who left within the last month – churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer (tenure), contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

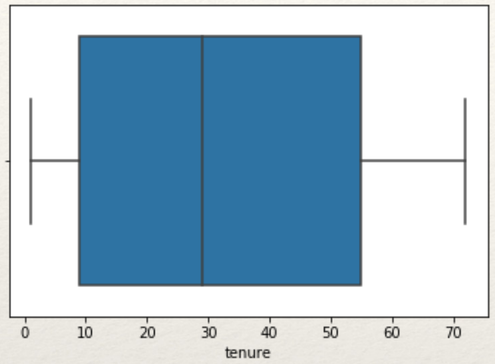
Exploratory Data Analysis (EDA)

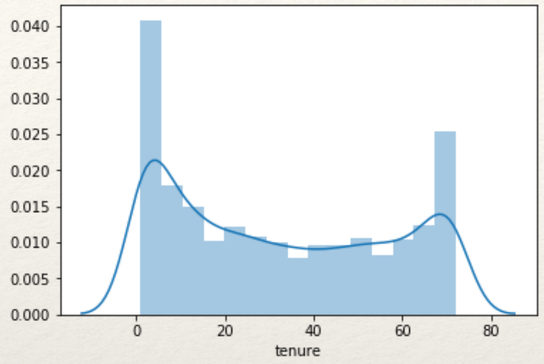
Drop missing data

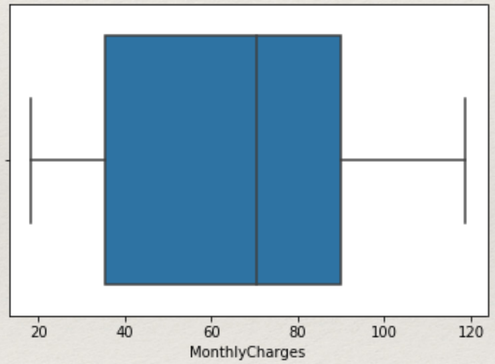
- As to variable *TotalCharges*:
- Transfer the data type “string” to a float value which it is supposed to be
- Drop data with null value
- Check other string values - no missing data
- Encode variables

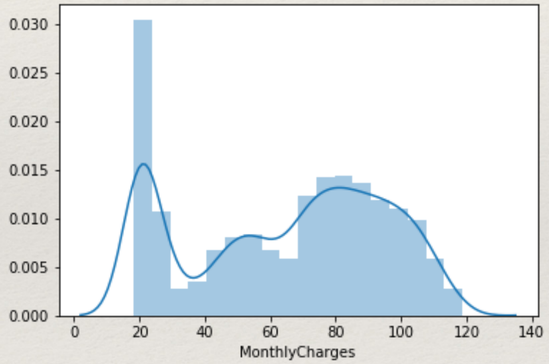
```
<class 'pandas.core.frame.DataFrame'>
Index: 7043 entries, 7590-VHVEG to 3186-AJIEK
Data columns (total 20 columns):
gender                7043 non-null object
SeniorCitizen         7043 non-null int64
Partner               7043 non-null object
Dependents            7043 non-null object
tenure                7043 non-null int64
PhoneService          7043 non-null object
MultipleLines         7043 non-null object
InternetService       7043 non-null object
OnlineSecurity        7043 non-null object
OnlineBackup          7043 non-null object
DeviceProtection      7043 non-null object
TechSupport           7043 non-null object
StreamingTV           7043 non-null object
StreamingMovies       7043 non-null object
Contract              7043 non-null object
PaperlessBilling      7043 non-null object
PaymentMethod         7043 non-null object
MonthlyCharges        7043 non-null float64
TotalCharges          7043 non-null object
Churn                 7043 non-null object
dtypes: float64(1), int64(2), object(17)
memory usage: 1.1+ MB
```

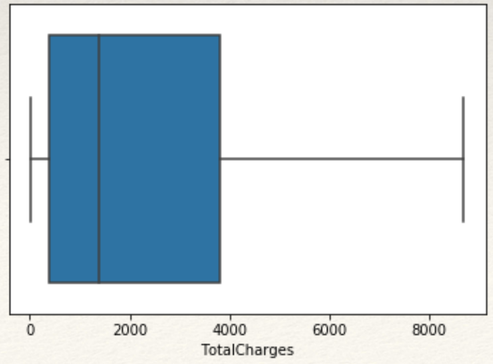
Outliers and Transformation (Univariate analysis & data visualization)

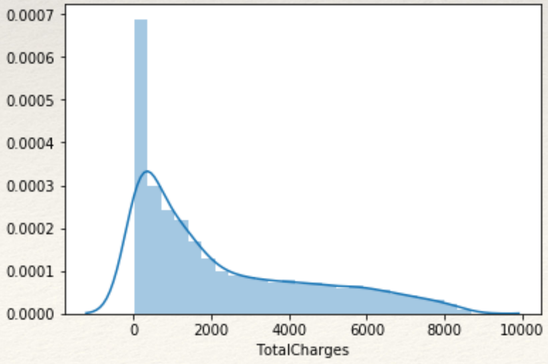




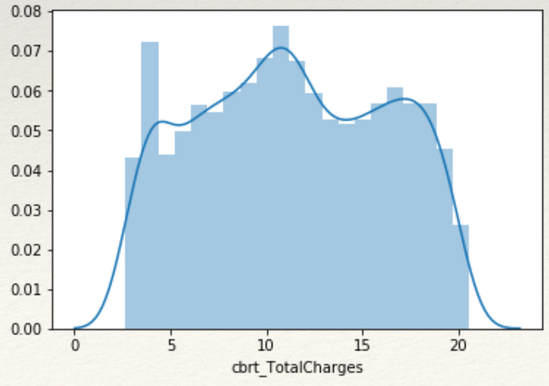






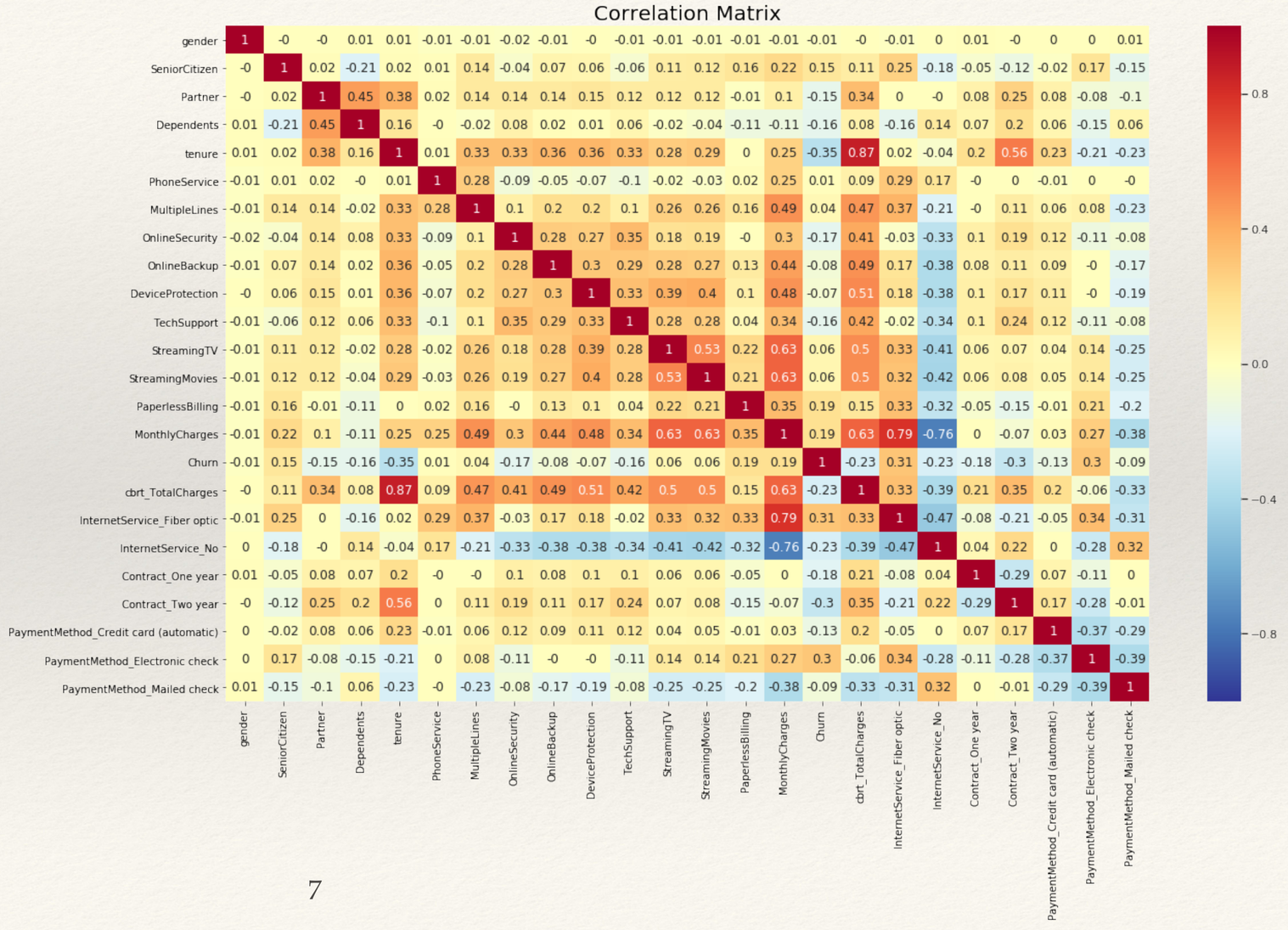


- As for *tenure*, *MonthlyCharges*, and *TotalCharges*:
- Under the criteria of 1.5 IQR, there are no outliers.
- Plot the distribution of these there variables.
- *TotalCharges* is highly right skewed and we apply cubic-root transformation on it.
- (cbrt_TotalCharges)



Underlying Multicollinearity

- Through the means of each variable, we find that over 90% of customer order phone service, which indicates its high correlation with constant.
- Correlation matrix
- We can see that there exists multicollinearity problems. The darker the color, the more severe the problem.



- Findings:
- "tenure" are strongly correlated with "cbrt_TotalCharges", because the longer a customer stays with the company, the more she/he is likely to spend.
- Internet service and monthly charges are strongly correlated, probably because the internet service fee is high.
- P.S. There are still variables that are mildly correlated and variables group (>2) can also create multicollinearity.
- Since higher correlations exist among some variables, We need variable selections to solve multicollinearity.

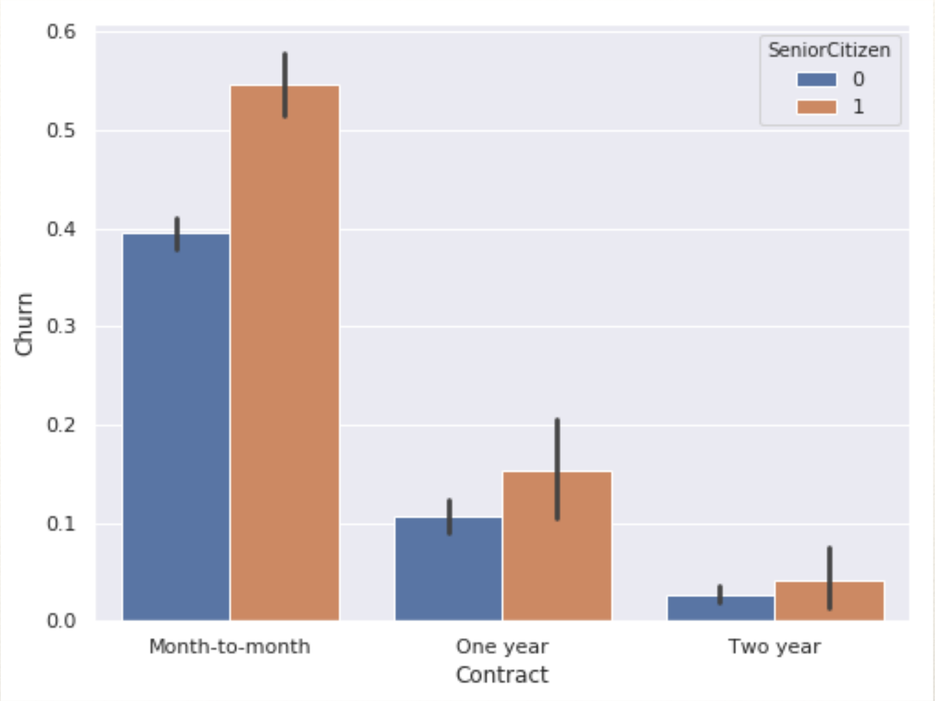
Bivariate analysis & data visualization

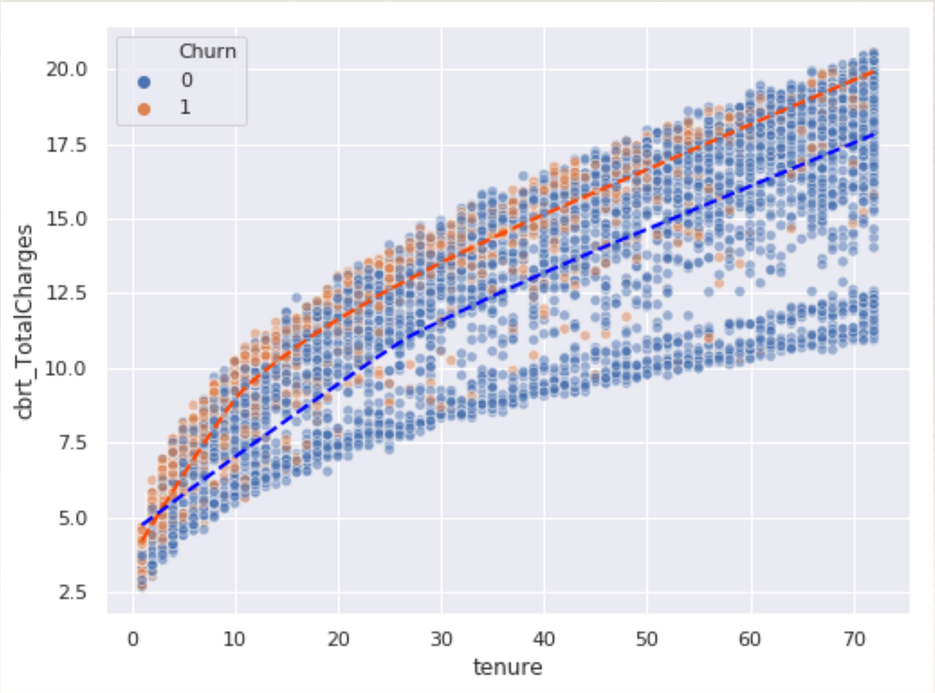
Let data speak!

Bivariate analysis & data visualization

- Longer contracts lead to lower churns.
- Senior citizens are more likely to churn.

- Conditional on tenure, the larger the total charges, the more likely for a customer to churn. (Demand Theory)
- The longer the tenure, the scarcer the orange points, the less likely for a customer to churn. (Loyalty)



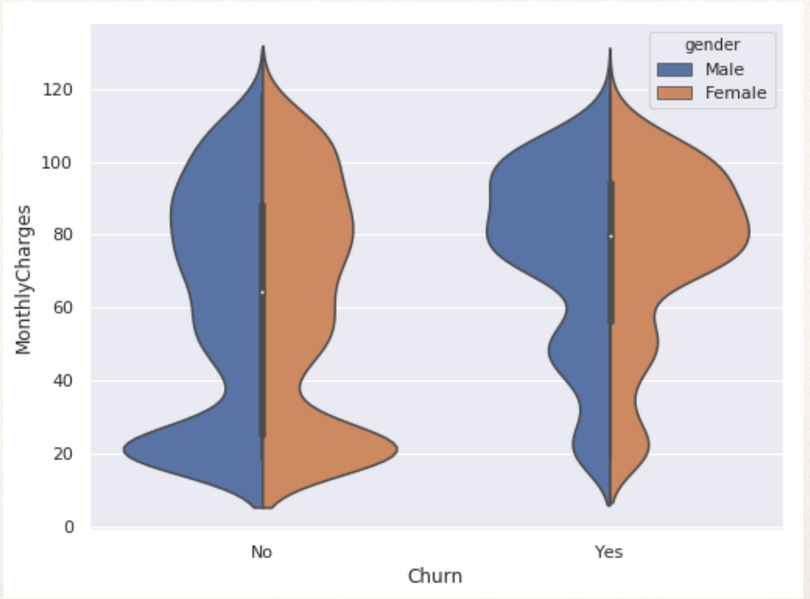


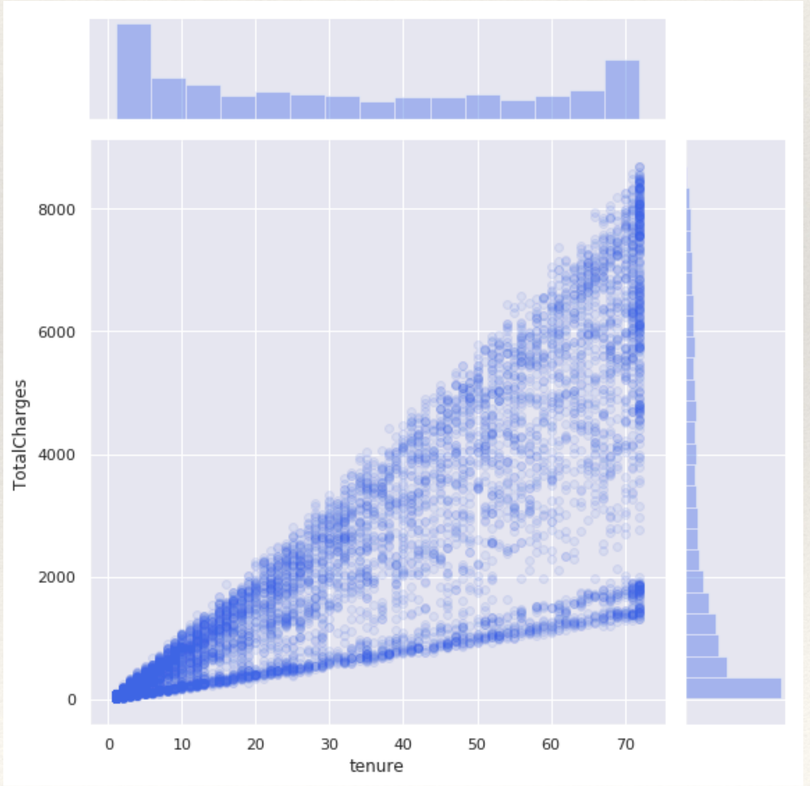
Let data speak!

Bivariate analysis & data visualization (continued)

- Customers who churns have higher MonthlyCharges. (Demand Theory)
- Gender seems to have no effect on MonthlyCharges. (Almost Symmetric)

- Though we assume customers generate the same revenue, it is not true in reality:
- The scatter plot tell us that it is not wise to retain every customer but customers whose monthly charges are high, since the difference of total charges accumulates.
- The histogram tells us that while most customers generate mild revenue, few customers generate huge revenue. We should focus on these minorities, who are called Most Valuable Customers.





Scoring Model - Logistic Regression

Assumptions Checking

- Binary logistic regression requires the dependent variable to be binary.
  Check.
- Logistic regression requires the observations to be independent of each other.
  There is no information on how this sample is acquired. Nevertheless, we need to assume this to continue our journey.
- Logistic regression requires there to be little or no multicollinearity among the independent variables.
  This assumption is actually violated. Later, we will perform feature selections to solve this problem, but then the result will be more difficult to interpret.
- Logistic regression assumes linearity of independent variables and log odds.
  This is an assumption that we cannot check.
- Logistic regression typically requires a large sample size. A general guideline is that a minimum of 10 observations is needed with the least frequent outcome for each independent variable in the model.
  $10 * 19 / 0.266 = 714 < 7032$. So, we have enough observations.

Benchmark Logistic Regression

Model fitting

Model:	Logit	Pseudo R-squared:	0.285
Dependent Variable:	Churn	AIC:	5874.2724
Date:	2018-12-23 02:26	BIC:	6038.8698
No. Observations:	7032	Log-Likelihood:	-2913.1
Df Model:	23	LL-Null:	-4071.7
Df Residuals:	7008	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	8.0000		

```
naive_logit = smf.logit("""Churn ~ gender
                        + SeniorCitizen
                        + Partner
                        + Dependents
                        + tenure
                        + PhoneService
                        + MultipleLines
                        + InternetService
                        + OnlineSecurity
                        + OnlineBackup
                        + DeviceProtection
                        + TechSupport
                        + StreamingTV
                        + StreamingMovies
                        + Contract
                        + PaperlessBilling
                        + PaymentMethod
                        + MonthlyCharges
                        + TotalCharges""", data = reg_data).fit()
```

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	1.1653	0.8151	1.4296	0.1528	−0.4324	2.7629
InternetService[T.Fiber optic]	1.7475	0.7981	2.1896	0.0286	0.1833	3.3117
InternetService[T.No]	−1.7863	0.8073	−2.2127	0.0269	−3.3685	−0.2041
Contract[T.One year]	−0.6608	0.1076	−6.1420	0.0000	−0.8717	−0.4499
Contract[T.Two year]	−1.3571	0.1765	−7.6907	0.0000	−1.7030	−1.0113
PaymentMethod[T.Credit card (automatic)]	−0.0878	0.1141	−0.7696	0.4416	−0.3114	0.1358
PaymentMethod[T.Electronic check]	0.3045	0.0945	3.2220	0.0013	0.1193	0.4897
PaymentMethod[T.Mailed check]	−0.0576	0.1149	−0.5011	0.6163	−0.2828	0.1676
gender	−0.0218	0.0648	−0.3369	0.7362	−0.1488	0.1052
SeniorCitizen	0.2168	0.0845	2.5644	0.0103	0.0511	0.3825
Partner	−0.0004	0.0778	−0.0049	0.9961	−0.1529	0.1522
Dependents	−0.1485	0.0897	−1.6548	0.0980	−0.3244	0.0274
tenure	−0.0606	0.0062	−9.7157	0.0000	−0.0728	−0.0484
PhoneService	0.1715	0.6487	0.2643	0.7915	−1.1000	1.4429
MultipleLines	0.4484	0.1773	2.5296	0.0114	0.1010	0.7958
OnlineSecurity	−0.2054	0.1787	−1.1496	0.2503	−0.5556	0.1448
OnlineBackup	0.0260	0.1754	0.1485	0.8820	−0.3177	0.3698
DeviceProtection	0.1474	0.1764	0.8356	0.4034	−0.1983	0.4931
TechSupport	−0.1805	0.1806	−0.9994	0.3176	−0.5345	0.1735
StreamingTV	0.5905	0.3263	1.8096	0.0704	−0.0491	1.2301
StreamingMovies	0.5993	0.3267	1.8345	0.0666	−0.0410	1.2396
PaperlessBilling	0.3424	0.0745	4.5956	0.0000	0.1963	0.4884
MonthlyCharges	−0.0403	0.0318	−1.2704	0.2039	−0.1026	0.0219
TotalCharges	0.0003	0.0001	4.6571	0.0000	0.0002	0.0005

Resampling method to assess the quality of the estimates:

Bootstrap Resampling

- Non-parametric confidence intervals

	[0.025	0.975]
Intercept	−0.287500	2.730001
InternetService[T.Fiber optic]	0.200491	3.382637
InternetService[T.No]	−3.439704	−0.216267
Contract[T.One year]	−0.849572	−0.454907
Contract[T.Two year]	−1.671539	−1.035965
PaymentMethod[T.Credit card (automatic)]	−0.313460	0.131301
PaymentMethod[T.Electronic check]	0.122052	0.488815
PaymentMethod[T.Mailed check]	−0.302195	0.159406
gender	−0.139827	0.120095
SeniorCitizen	0.044571	0.362554
Partner	−0.135465	0.145475
Dependents	−0.331185	0.013122
tenure	−0.076620	−0.048057
PhoneService	−1.107583	1.472055
MultipleLines	0.126836	0.792861
OnlineSecurity	−0.549326	0.143191
OnlineBackup	−0.339743	0.369838
DeviceProtection	−0.229827	0.491925
TechSupport	−0.529255	0.157508
StreamingTV	−0.026951	1.264425
StreamingMovies	−0.061256	1.240612
PaperlessBilling	0.202718	0.492601
MonthlyCharges	−0.106536	0.019915
TotalCharges	0.000177	0.000501

Comparison of Parametric and Non-parametric Confidence Intervals

- In theory, the coefficient of maximum likelihood estimation tends to normal distribution so that the parametric confidence interval and non-parametric confidence interval are asymptotically the same. Since this sample is large, they are almost the same.

Average Percentage Width Difference

```
In [36]: ((confint_df["0.975]"]-confint_df["[0.025"]) / (benchmark_logit.conf_int()[1]-benchmark_logit.conf_int()[0])).mean() - 1
```

```
0.004441360505204939
```

Average Percentage Left End Difference

```
In [37]: ((confint_df["[0.025"]-benchmark_logit.conf_int()[0])/(benchmark_logit.conf_int()[1]-benchmark_logit.conf_int()[0])).mea
```

```
-0.004169868607085903
```

Average Percentage Right End Difference

```
In [38]: ((confint_df["0.975]"]-benchmark_logit.conf_int()[1])/(benchmark_logit.conf_int()[1]-benchmark_logit.conf_int()[0])).mea
```

```
0.0002714918981191374
```

Here, we did not apply any modification on data.

- The effect of the length of contract is significant, which is intuitively correct.
- Senior citizens are more likely to churn, probably because they have low income and are more price sensitive.
- People who stay longer with the company are less likely to churn, probably because they are more satisfied with the service.
- People who have multiple lines are more likely to churn, probably because customers with multiple lines care more about about phone service.
- Conditional on tenure, people who have bigger total charges are more likely to churn. This is compatible with demand theory: the higher the price, the more likely for a customer to leave.

However, there are some problems here.

- As we mentioned before, Intercept is strongly correlated with PhoneService since PhoneService is almost a constant. So here both Intercept and PhoneService are insignificant.
- It is interesting that people with no internet service are less likely to churn, while people with more advanced internet service such as fiber optic are more likely to churn. This is strange because if a person does not order internet service, the only link between she/he with the company is phone service. A possible explanation is endogeneity. For example, people who choose more advanced internet access may have more information for comparing services from different companies, so they are more likely to churn.
- The fact that the method of paperless billing boosts churn probability is hard to explain because it is neither customers' behavior nor their attribute. Maybe it is because people who have internet access will choose paperless billing, which is faster and more convenient, and as mentioned before, having access to internet increases the probability of churning.

In the sense of economic significance:

- Contract is economically significant. One-year contracts push down log-odds of churn probability by 0.6608 and two-year contracts push that down by 1.3571.

In conclusion:

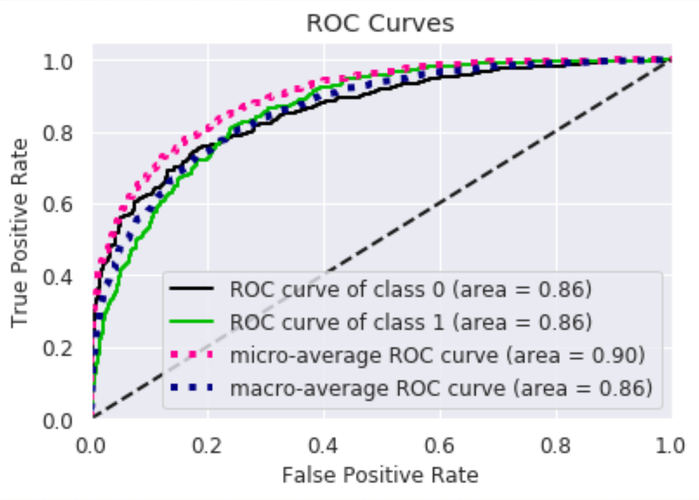
- It is always a good idea to sign a long-term contract with customers to prevent churns.
- We need to focus more on new customers because: 1) They are more likely to churn. 2) The business relationship becomes gradually solid as time goes by. (customer cultivation)
- It is less desired to acquire senior citizen customers, but once they are acquired, we need to pay special attention to them.
- Since price margin is not available, whether decreasing price is a good idea requires further analysis that is not possible here.
- Be careful with customers who have multiple lines, they probably pay more but also leave easily.

Logistic Regression with Feature Selections and Transformations

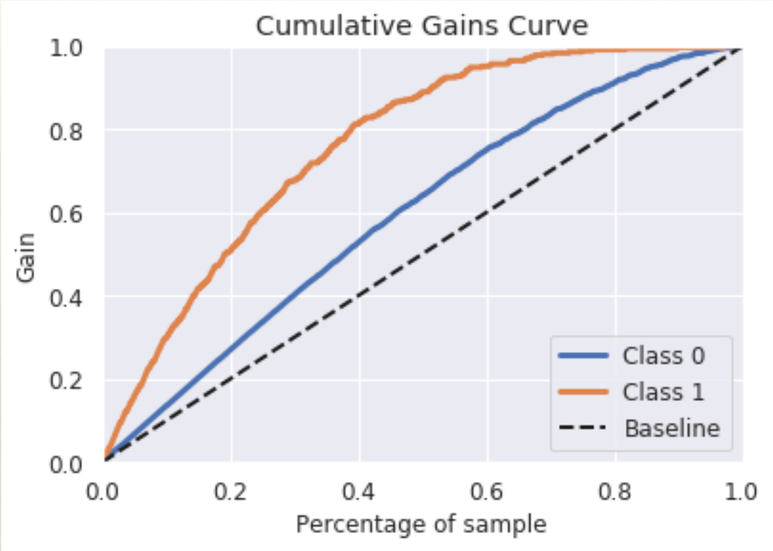
- We cannot compare different models on training set, or else the model with the most variables almost always wins. This is because models with more variables tend to absorb more noise from the training set. Here we apply the rule of a training : test = 80 : 20 split and compare three logistic models:
- **Benchmark Logistic Model (with all variables in their original form)**
- **Logistic model with variable (cubic-root) transformations**
- **Logistic model with variable transformations and variable selections (according to BIC)**

- P.S. Strictly speaking, we need to divide the original sample into three parts: training : dev : test = 6 : 2 : 2.
- Train models on training set.
- Select models and cutoff values on dev set.
- Predict on test set.
- Here, we omit dev set for simplicity purposes.

Benchmark Logistic Model

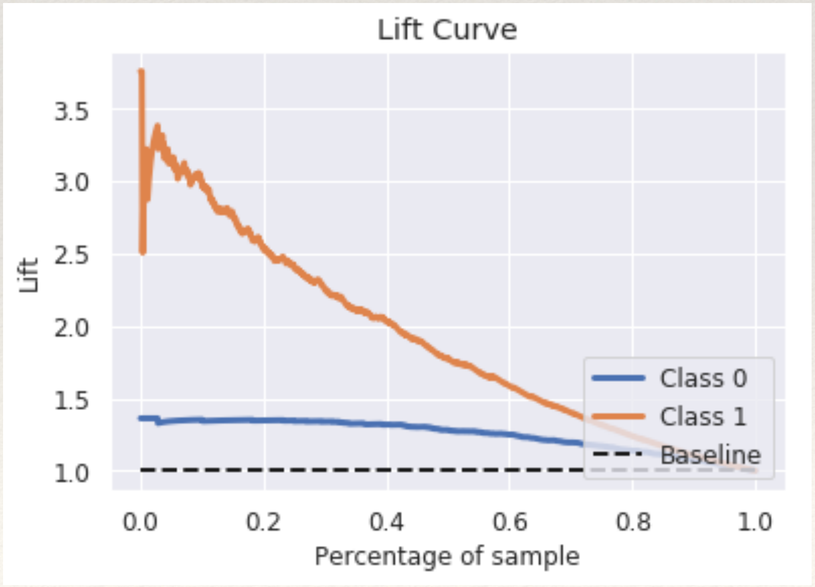


- A receiver operating characteristic curve, i.e., ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied (when the threshold decreases, true positive rate increases and so does false positive rate). The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups. Usually an AUC over 0.7 is a good prediction. Therefore, we have a good model here.



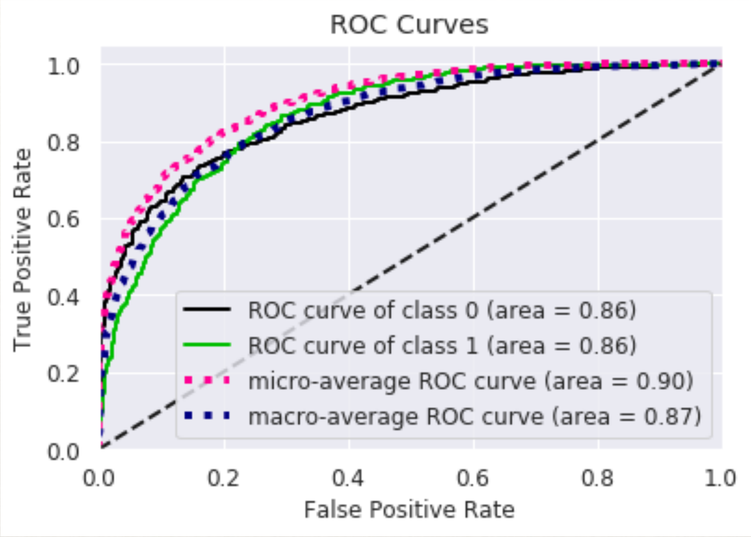
- The cumulative gains chart shows the percentage of the overall number of cases in a given category "gained" by targeting a percentage of the total number of cases. For example, the orange line crosses the point (0.4, 0.8), which means if the company contacts 40% of the customer, 80% of them actually will churn. Again, the result is satisfactory.

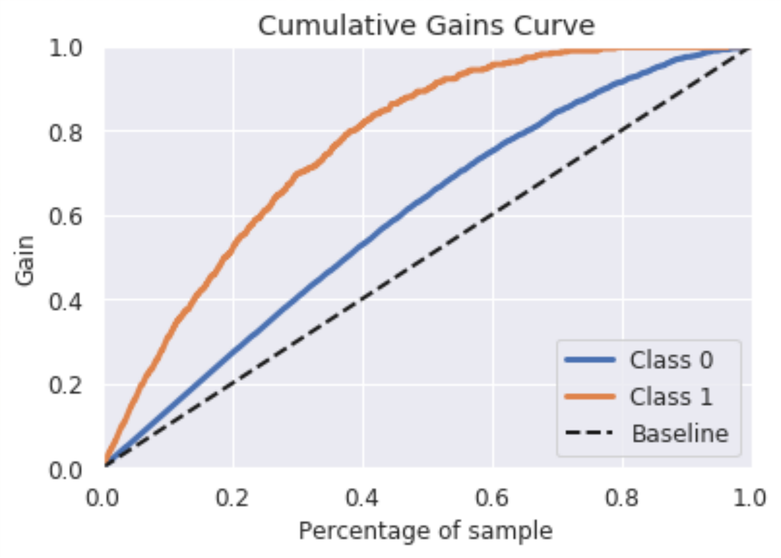
- The lift chart is derived from the cumulative gains chart; the values on the y axis correspond to the ratio of the cumulative gain for each curve to the baseline. For example, the orange line crosses the point (0.4, 2) which means if the company contacts 40% of the customer, the ratio of cumulative gain to the base line is 2. The value of baseline here is 0.4, so the cumulative gain at this point is 0.8, which is compatible to the result above.

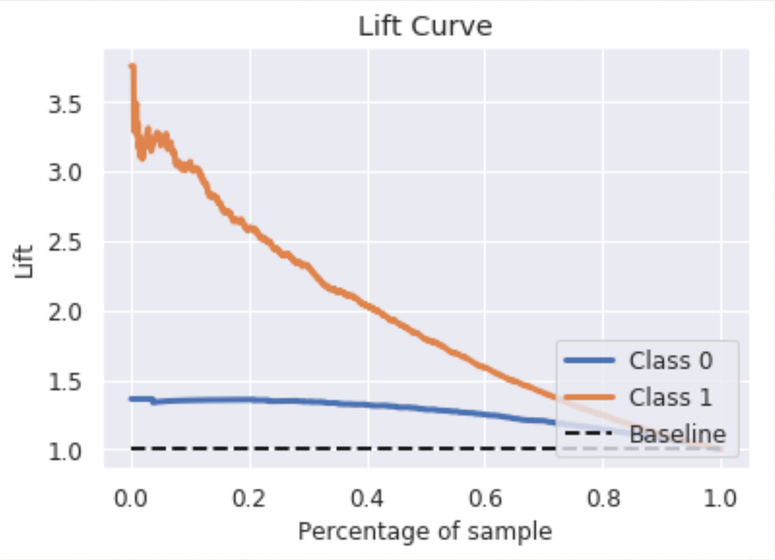


Logistic Model with Variable Transformations

- Everything is good as before.

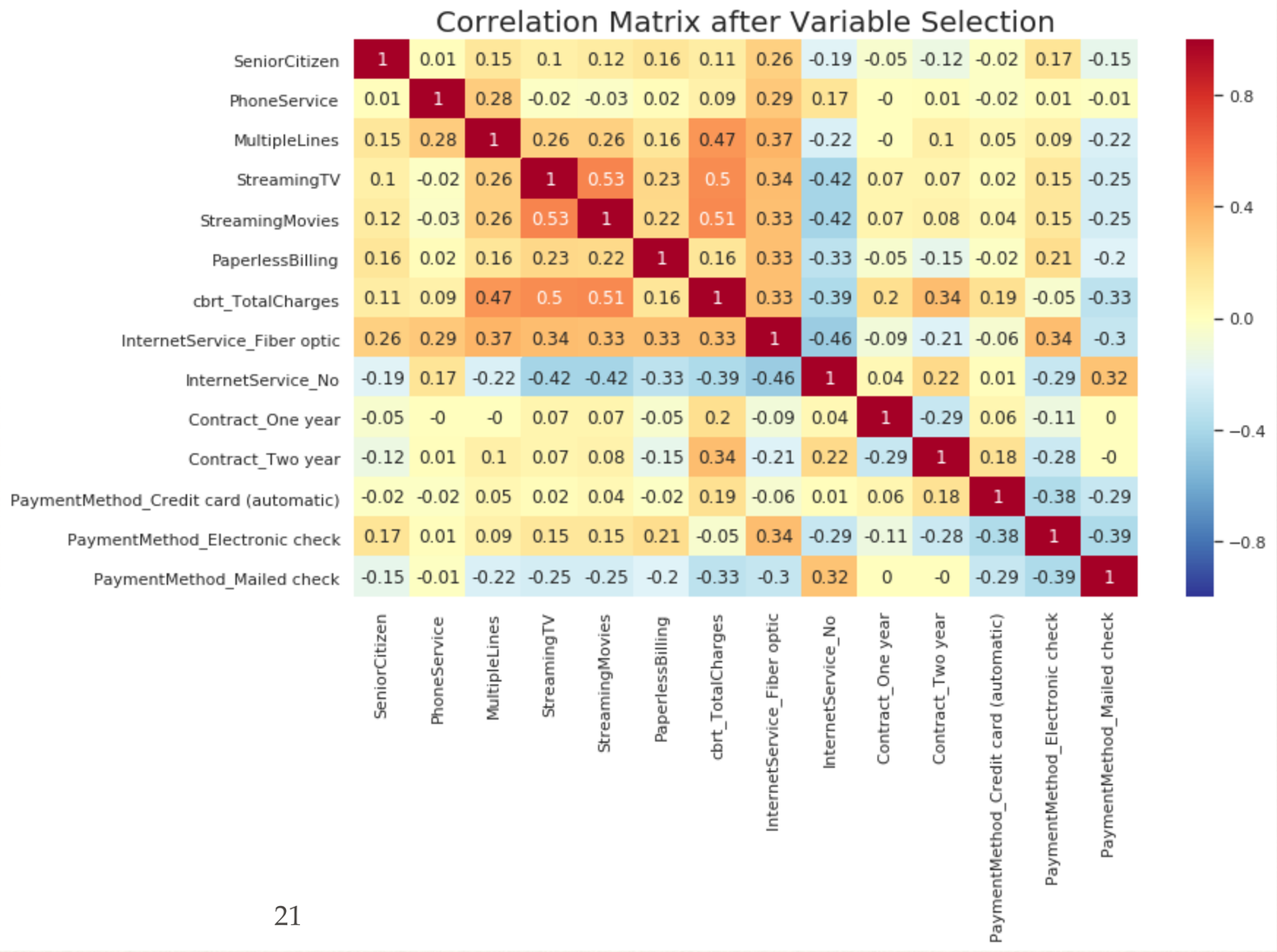






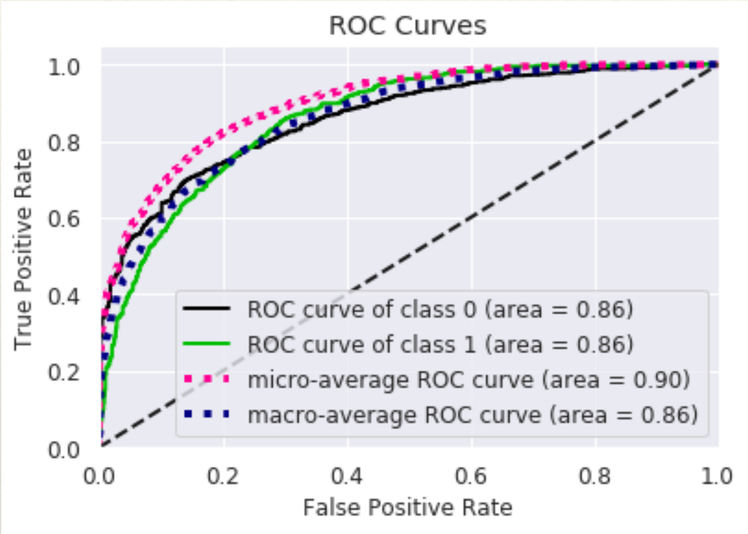
Logistic Model with Variable Transformations and Variable Selections

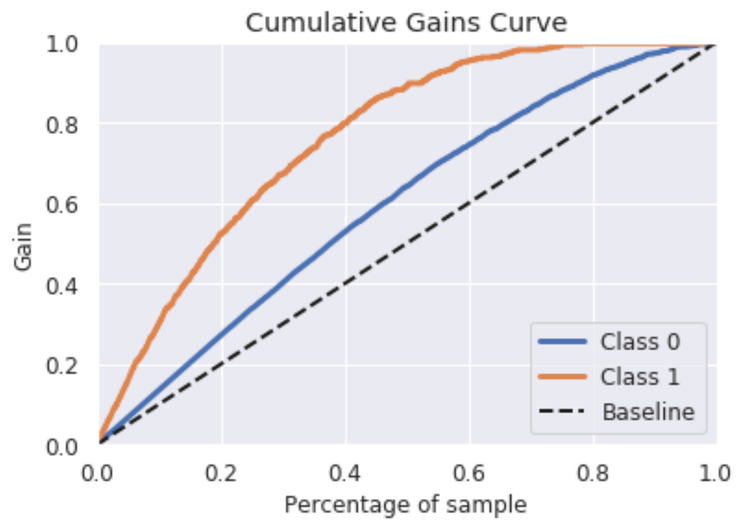
- Here we apply Backward Stepwise Selection Method.
- We see that the BIC of this model is the lowest among the three.
- After variable selection, the multicollinearity problem is alleviated. We meet another assumption!

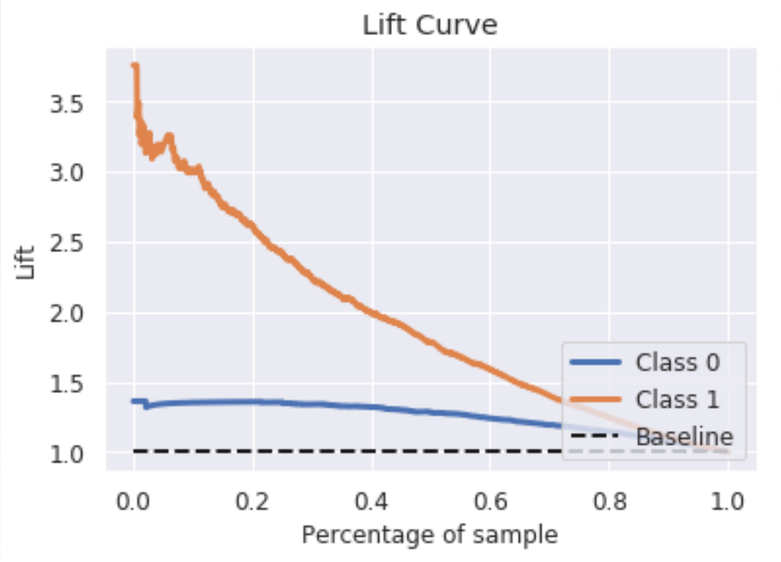


Logistic Model with Variable Transformations and Variable Selections (continued)

- Good as always.



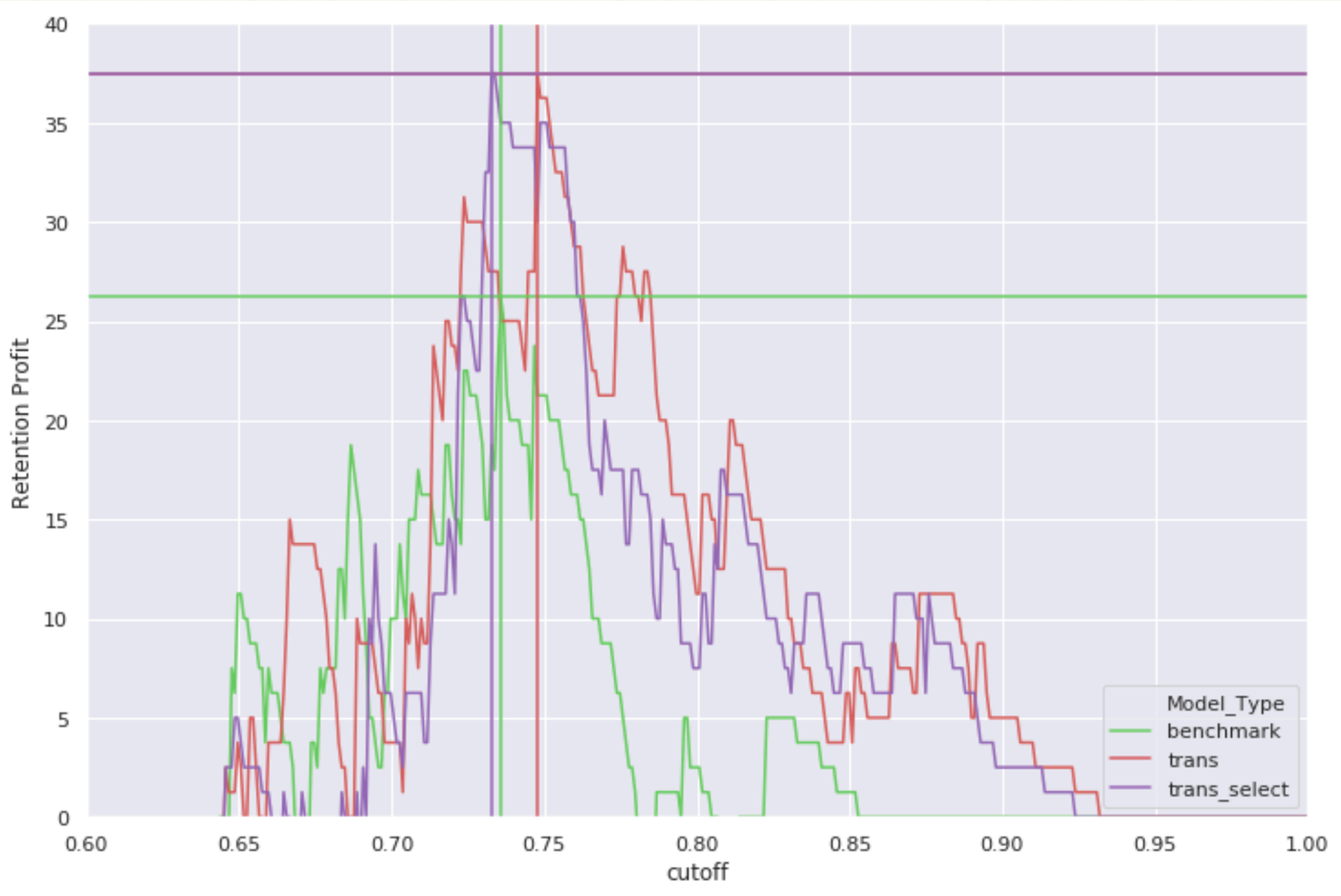




Conclusions - Business Case Simulations

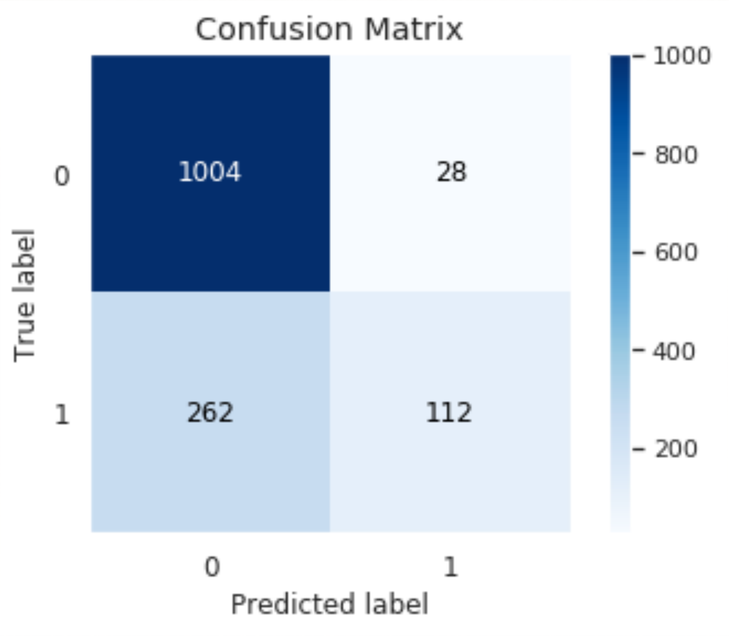
Comparisons of the Three Models

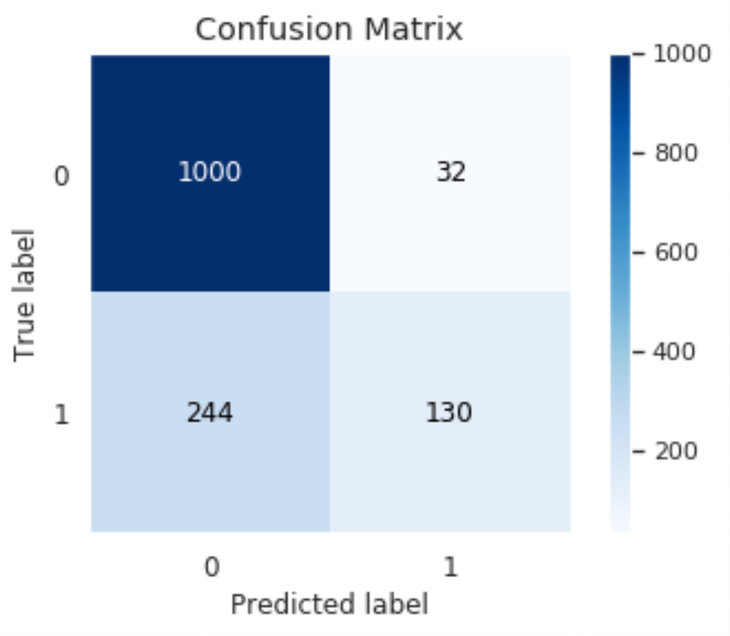
- Retention Profit VS Cutoff
- We have seen ROC, gain and lift plots of the three models. Clearly, they are hard to differentiate. Therefore, wee need to plot the cutoff and profit, which is more related to the final objective.
- The lowest cutoff rate when retention profit hits zero doesn't indicate the best model because we have not taken into account the difference in model prediction power. However, if the company wants to maximize the retention profit, here both the non-benchmark models are potential alternatives.

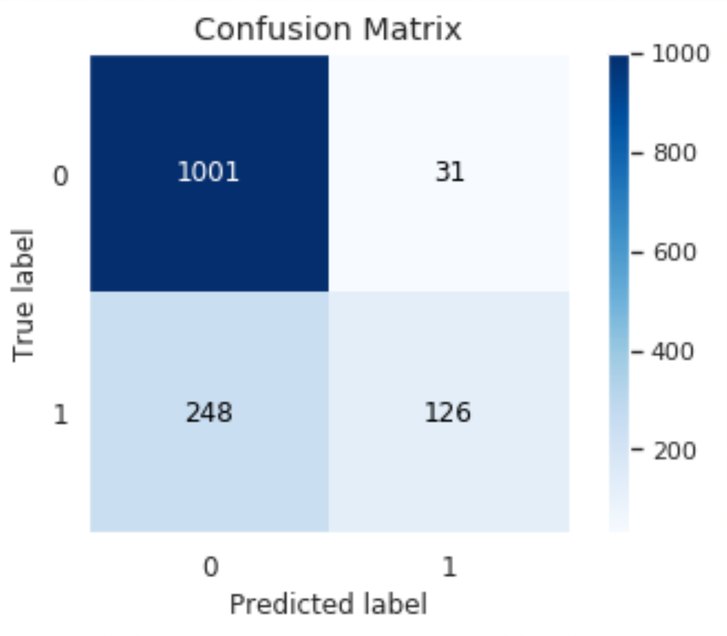


The Optimal Model for Performance

- From the test sample, we see that under the assumption of Break-Even Budget, logistic regression with variable transformations can acquire the biggest number of customers who will churn, which is 130!







- BRONZE: Benchmark Logistic Regression
- GOLD: Logistic Regression with Variable Transformations
- SILVER: Logistic Regression with Variable Transformations and Selections

Cutoff Selection and Economical Parameters

- **Cutoff:**
- **Number of Customers to Call in the Original Sample:**
- If the Telco Sector is to contact a new group of customers as large as the original sample size (after dropping 11 missing data), this is the target size (in term of number of customers) to contact through the Retention Commercial Campaign in order to have a Break-even Balance.
- **Churn Rate:**
- **Rate of Successful Retention:**
- **Expected Number of Customers Retained through Campaign:**

	Value
Cutoff:	0.6456
Number of Customers to Call in the Original Sample:	810.2304
Churn Rate:	0.2658
Rate of Successful Retention:	0.2006
Expected Number of Customers Retained through Campaign:	162.5462

Conclusions

- We use the benchmark logistic model (model with all original variables) to get the insight of customer churns, from which we conclude:
  1) Signing long-term contracts is the best strategy.
  2) Special attention needs to be paid to new customers (who churns easily) to transform convert them into long tenured customers (who rarely churns).
  3) Customers with multiple lines may care more about service and leave easily.
  4) Senior customers churns easily, probably due to their higher sensitivity to price.
  5) The feasibility of pricing strategy requires further analysis.
- We apply the logistic model with variable transformations and variable selections to business simulation to get the best predictive performance, from which we conclude:
  1) The cutoff rate is about 0.646, which means if a customer's predicted churn probability is higher than this value, she/he needs to be contacted.
  2) The number of customers to call in the origin sample (of which the size is 7032 observations after dropping 11 missing values) is about 810.
  3) The churn rate is about 0.266.
  4) The expected number of customers retained through this campaign is about 163.
  5) The rate of successful retention (customers retained / customers contacted) contacted is 0.2.

Limitations

- We confine ourselves to logistic regression models as required by the project guideline, while there may be other supervised learning models which provide better predictions.
- We assume each retained customer generate the same revenue, which is usually not true as we have shown in previous exploratory data analysis: Very few customers generate extremely high revenue, who are called Most Valuable Customers and require higher attention.

References

- Songul Albayrak: Customer churn prediction in telecommunication
- Clement Kirui, Li Hong, Wilson Cheruiyot, Hillary Kirui: Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining, *International Journal of Computer Science Issues*
- Wikipedia: Sensitivity and specificity https://en.wikipedia.org/wiki/Sensitivity_and_specificity
- Statistics Solutions: Assumptions of Logistic Regression https://www.statisticssolutions.com/assumptions-of-logistic-regression/
- IBM Knowledge Center: Cumulative Gains and Lift Charts https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/tutorials/mlp_bankloan_outputtype_02.html