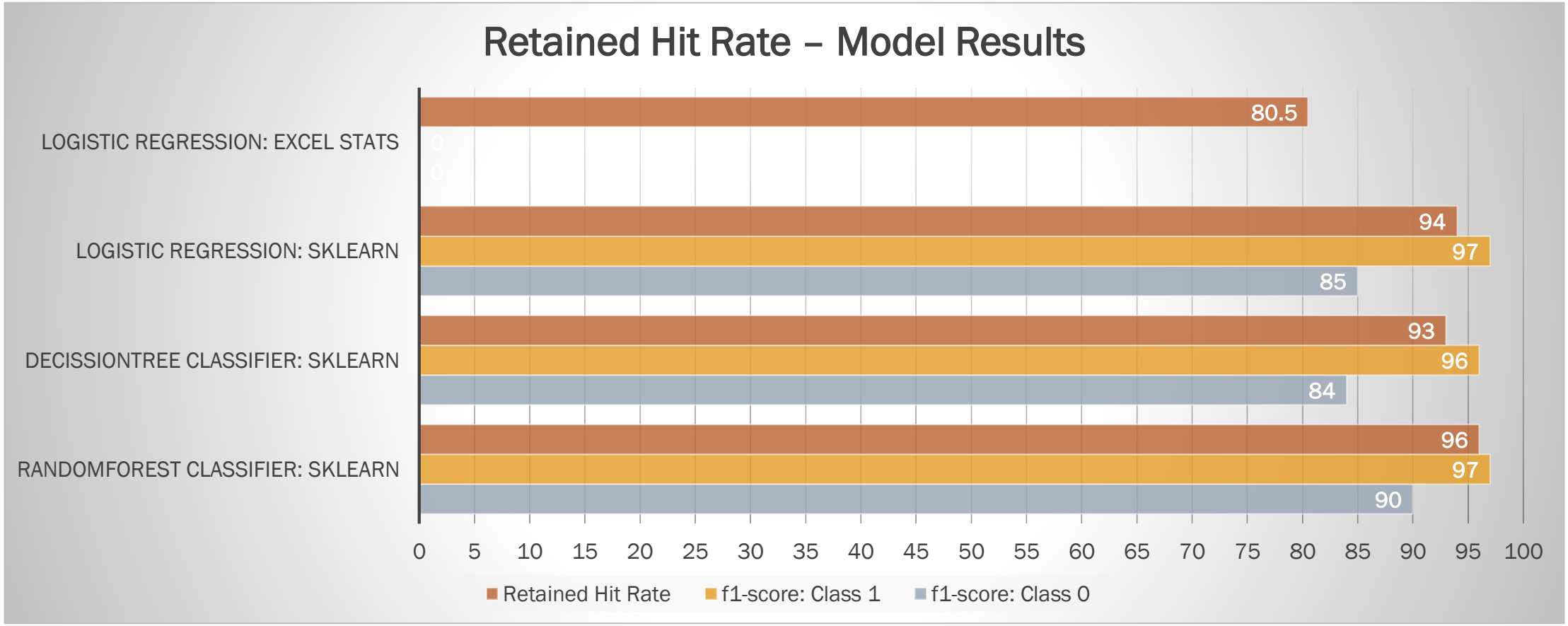
Retail Relay

Customer Churn Prediction

- RYAN TIMBROOK
- STEVE SKEELS

Model Performance Results



Exploratory Data Analysis

- Created new variables from existing data set:
  - ‘days1’ – Delta between ‘created’ & ‘firstorder’
  - ‘days2’ – Delta between ‘firstorder’ & ‘lastorder’
  - ‘weekday’ – Dummy variable to differentiate between ‘faveday’ on a weekday or weekend
  - Created dummy variables for each day of the week and each city.
- Transformed data in set:
  - Corrupted/missing date information
  - Transformed ratio data to percentages

custid	Computer generated ID to identify customers throughout the database
retained	1, if customer is assumed to be active, 0 = otherwise
created	Date when the contact was created in the database - when the customer joined
firstorder	Date when the customer placed first order
lastorder	Date when the customer placed last order
esent	Number of emails sent
eopenrate	Number of emails opened divided by number of emails sent
eclickrate	Number of emails clicked divided by number of emails sent
avgorder	Average order size for the customer
ordfreq	Number of orders divided by customer tenure
paperless	1 if customer subscribed for paperless communication (only online)
refill	1 if customer subscribed for automatic refill
doorstep	1 if customer subscribed for doorstep delivery
train	1 if customer is in the training database
favday	Customer's favorite delivery day
city	City where the customer resides in
openrate	"eopenrate" converted to percentage
clickrate	"eclickrate" converted to percentage
days1	# of days between account creation and first order
days2	# of days between first order and last order
Monday, Tuesday, Weds…	Dummy variables created from "favday"
citycho, citydcx, cityric	Dummy variables created from "city"
weekday	Dummy variable created from "favday" 1 if weekday, 0 if weekend

Correlation

	retained	*Monday*	*Tuesday*	*Wednesday*	*Thursday*	*Friday*	*Saturday*	*Weekday*	*clickrate*	*esent*	*openrate*	*days1*	*days2*	*avgorder*	*ordfreq*	*paperless*	*refill*	*doorstep*	*citycho*	*citydcx*	*cityric*
retained	1																				
Monday	0.0162617	1																			
Tuesday	-0.010114	-0.281041	1																		
Wednesday	-0.011553	-0.229624	-0.22867	1																	
Thursday	-0.004181	-0.241233	-0.24023	-0.19628	1																
Friday	0.0145846	-0.241753	-0.240748	-0.196704	-0.206648	1															
Saturday	0.0240926	-0.109837	-0.109381	-0.08937	-0.093888	-0.09409	1														
Weekday	0.0094784	0.1357434	0.1351793	0.1104483	0.116032	0.1162821	-0.809153	1													
clickrate	0.049195	0.0172811	0.0009064	-0.010598	-0.005689	0.0029299	0.0217891	0.0109953	1												
esent	0.7194795	0.0286597	0.0014848	-0.015659	0.0067639	-0.016627	0.0168135	0.0127688	-0.093347	1											
openrate	0.0753256	0.0229724	0.0118675	-0.014024	-0.0102	-0.006933	0.0234657	0.0119303	0.5547395	-0.108485	1										
days1	-0.017506	0.0219506	0.0078081	0.0056315	-0.013471	0.0015783	-0.021881	0.0412592	-0.006642	0.0698547	-0.02244	1									
days2	0.0077427	0.0552412	0.0038264	-0.012011	-0.008306	-0.013481	-0.055869	0.0495496	0.0173522	0.216823	0.0321765	0.1166949	1								
avgorder	0.0036069	0.030093	-0.011445	-0.002652	-0.005873	-0.033785	0.0063355	-0.034089	-0.029642	0.1109776	-0.028172	0.0812222	0.1934412	1							
ordfreq	0.0080654	0.0137635	0.0004658	0.0032069	-0.007299	-0.012384	-0.015114	-0.001473	0.060089	0.0379748	0.0372906	0.0277608	0.0249292	0.0551332	1						
paperless	0.1771657	0.0367687	0.0173775	-0.037296	-0.029401	0.0271021	0.0563559	0.0332793	0.2047765	0.0124862	0.2421812	-0.230999	-0.21939	-0.14462	-0.02635	1					
refill	0.1028009	0.0241618	-0.012947	-0.010953	-0.005926	0.0271411	-0.01156	0.0361498	0.1428128	0.0493672	0.1357507	-0.00701	-0.003819	-0.057436	0.0643672	0.1772467	1				
doorstep	0.0688041	0.0340902	-0.028632	-0.007173	-0.014198	0.0066589	0.0260105	-0.013165	0.1005656	0.0443188	0.1044951	-0.019229	0.0211739	0.0484016	0.0929819	0.0983071	0.203284	1			
citycho	-0.089346	0.0735788	-0.016654	0.0474227	-0.004873	-0.005411	-0.128317	0.1542453	-0.061892	-0.044454	-0.082173	0.1238597	0.2265713	0.0672636	0.0261535	-0.230649	-0.019907	-0.103403	1		
citydcx	0.057577	-0.04635	-0.005782	-0.015534	-0.016967	-0.058741	0.1573302	-0.232567	0.1203516	-0.004259	0.1112398	-0.083809	-0.160267	-0.048713	-0.001398	0.2139064	-0.046408	0.2024265	-0.397157	1	
cityric	0.0072216	0.0174709	0.0053301	0.0081063	-0.012894	0.1052655	-0.160444	0.1968834	-0.077777	0.0364896	-0.052935	-0.021216	-0.032116	-0.025524	-0.0269	-0.029806	0.0686975	-0.106894	-0.490143	-0.48794	1

Regression

Models:

- After analyzing correlation analysis ran multiple different regression models.
- Prediction1 was the initial test of variables based on logic and correlation data.
- Prediction2 was the model that yielded best results.

Results:

Adjusted R S	0.57991286	
Prediction1	*Coefficients*	*P-value*
Intercept	0.2220628	0
esent	0.01855141	0
openrate	0.18670975	3.615E-212
days1	-9.099E-05	2.2218E-11
days2	-0.0002402	2.194E-194
avgorder	-0.0003282	1.3955E-15
ordfreq	-0.0753074	2.1046E-06
paperless	0.07446263	2.2784E-81
refill	0.04007095	4.6932E-12

Adjusted R S	0.58126179	
Prediction2	*Coefficients*	*P-value*
Intercept	0.23119746	1.594E-116
Weekday	0.00576404	0.44962238
clickrate	0.13756719	1.1507E-12
esent	0.01861797	0
openrate	0.16016458	1.403E-116
days1	-9.684E-05	1.1916E-12
days2	-0.0002468	8.644E-196
avgorder	-0.0003382	2.1178E-16
ordfreq	-0.0862907	6.0045E-08
paperless	0.07178251	1.593E-71
refill	0.03553837	2.5241E-09
doorstep	0.02338051	0.01026045
citycho	-0.0046133	0.58230561
citydcx	-0.0150343	0.05573477
cityric	-0.0232402	0.00410998

Model Interpretation

Predicted Retention & Hit Rate:

- 80.16% is highest hit rate achieved with Logistic Regression.
- Interesting results when only using ‘esent’ in a prediction model.
- We can do better; back to the drawing board…

Prediction1

a+bx	exp(a+bx)	p(sale)	Predicted	observed
3.44757406	31.424067	0.96915871	1	1

		Observed	
		0	1
Predicted	0	42	1
	1	1244	4932

Hit Rate: 79.98%

Prediction2

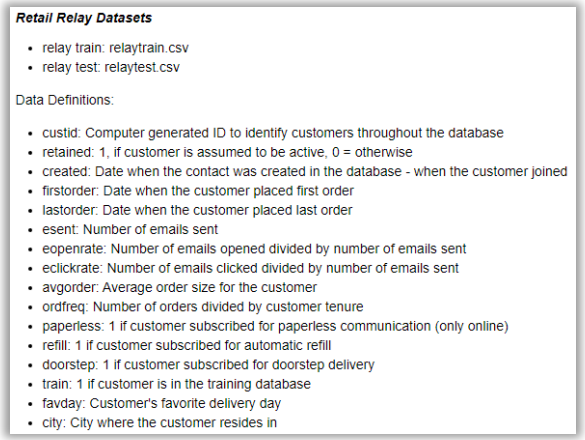
A+bx	exp(a+bx)	p(sale)	Predicted	Observed
3.44218825	31.2552777	0.96899732	1	1

		Observed	
		0	1
Predicted	0	52	0
	1	1234	4933

Hit Rate 80.16%

Data Collection & Preparation

- Relay Training Dataset Shape:
  - Rows: **24,578**
  - Columns: **16**
- Relay Training Dataset Shape:
  - Rows: **6,219**
  - Columns: **16**



Retail Relay Datasets

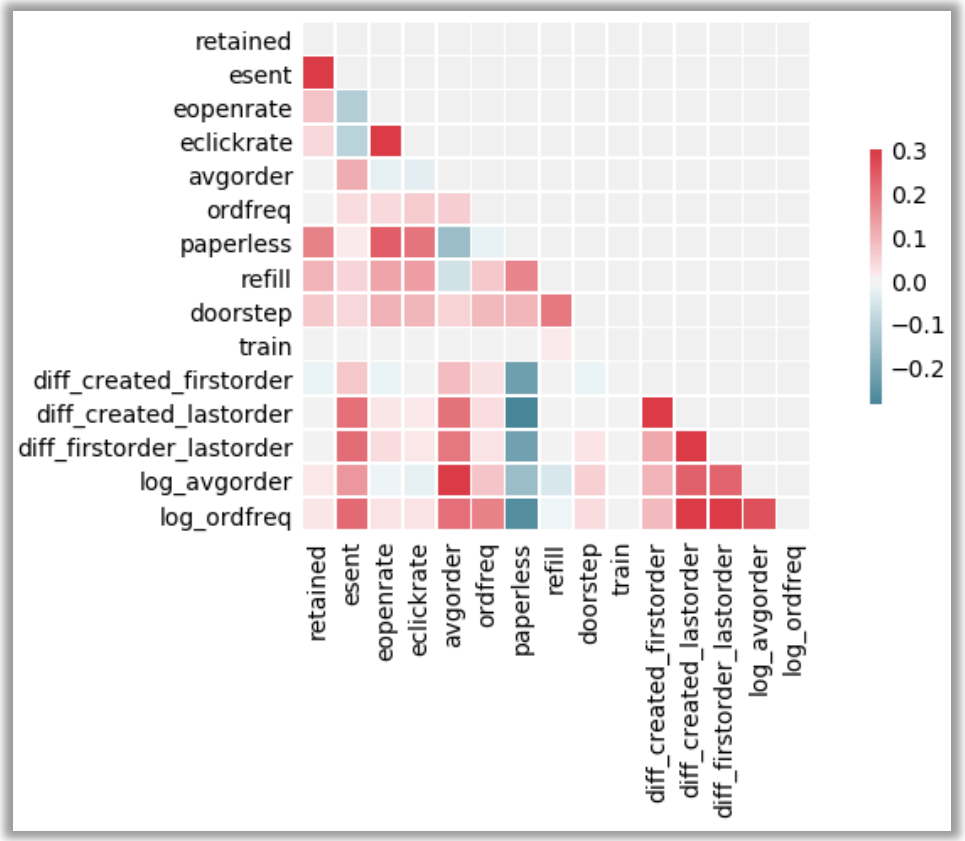
- relay train: relaytrain.csv
- relay test: relaytest.csv

Data Definitions:

- custid: Computer generated ID to identify customers throughout the database
- retained: 1, if customer is assumed to be active, 0 = otherwise
- created: Date when the contact was created in the database - when the customer joined
- firstorder: Date when the customer placed first order
- lastorder: Date when the customer placed last order
- esent: Number of emails sent
- eopenrate: Number of emails opened divided by number of emails sent
- eclickrate: Number of emails clicked divided by number of emails sent
- avgorder: Average order size for the customer
- ordfreq: Number of orders divided by customer tenure
- paperless: 1 if customer subscribed for paperless communication (only online)
- refill: 1 if customer subscribed for automatic refill
- doorstep: 1 if customer subscribed for doorstep delivery
- train: 1 if customer is in the training database
- favday: Customer's favorite delivery day
- city: City where the customer resides in

- Transformation Steps Taken:
  - Datasets **merged** into one **master dataset** for exploration, cleaning and transformation steps
    - Split to training/test prior to building models (80/20)
  - Datasets **cleaned** of N/A - **~600 total records** found to have bad data and were removed
    - Included bogus date values, missing order frequencies and average order values
  - Transformed **favday** and **city** into **category datatypes** for ML algorithms
  - Added '**days between**' attributes
    - **diff_created_firstorder**
    - **diff_firstorder_lastorder**
    - **diff_created_lasttorder**
  - Added **log normalization** attributes
    - **log_avgorder**
    - **log_ordfreq**

Exploratory Data Analysis
- Correlations

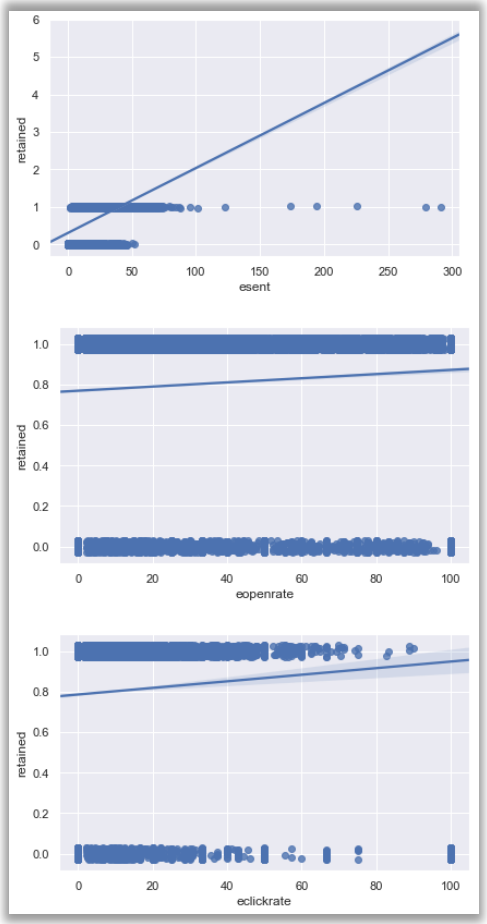


- **Emails sent** to customers has the **highest** positive **correlation** with retention.
- **Email open rate** and **average orders** made by a customer were shown to be the **second most import** features when **predicting 'retained'.**

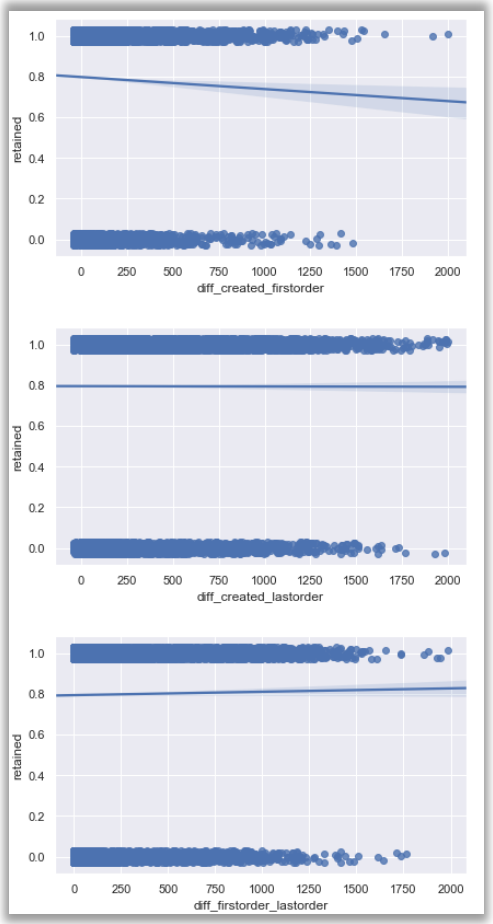
	retained	esent	eopenrate	eclickrate	avgorder	ordfreq	paperless	refill	doorstep	train	diff_created_firstorder	diff_created_lastorder	diff_firstorder_lastorder	log_avgorder	log_ordfreq
retained	1.000000	0.717509	0.075023	0.042477	0.004101	0.010744	0.180399	0.103513	0.066517	0.001237	-0.018451	-0.001032	0.009165	0.019206	0.019391
esent	0.717509	1.000000	-0.108275	-0.095311	0.114460	0.036206	0.013800	0.048110	0.042625	-0.002020	0.068500	0.212278	0.217064	0.147774	0.221813
eopenrate	0.075023	-0.108275	1.000000	0.553492	-0.025231	0.039097	0.238592	0.128667	0.104112	0.000652	-0.017952	0.021469	0.035987	-0.015881	0.023881
eclickrate	0.042477	-0.095311	0.553492	1.000000	-0.030405	0.061776	0.204426	0.136177	0.098179	0.001875	0.000341	0.015295	0.018232	-0.025602	0.026001
avgorder	0.004101	0.114460	-0.025231	-0.030405	1.000000	0.060279	-0.149162	-0.059884	0.049083	0.001493	0.087089	0.207144	0.200393	0.809817	0.212719
ordfreq	0.010744	0.036206	0.039097	0.061776	0.060279	1.000000	-0.022794	0.065581	0.090258	0.005865	0.029215	0.035061	0.025753	0.072540	0.180555
paperless	0.180399	0.013800	0.238592	0.204426	-0.149162	-0.022794	1.000000	0.177073	0.098609	0.000076	-0.222208	-0.284345	-0.217167	-0.149069	-0.265753
refill	0.103513	0.048110	0.128667	0.136177	-0.059884	0.065581	0.177073	1.000000	0.196226	0.013805	-0.007739	-0.007707	-0.004918	-0.049000	-0.014872
doorstep	0.066517	0.042625	0.104112	0.098179	0.049083	0.090258	0.098609	0.196226	1.000000	0.002799	-0.017288	0.010014	0.021815	0.053648	0.035646
train	0.001237	-0.002020	0.000652	0.001875	0.001493	0.005865	0.000076	0.013805	0.002799	1.000000	0.005383	0.004021	0.001807	0.002455	-0.003347
diff_created_firstorder	-0.018451	0.068500	-0.017952	0.000341	0.087089	0.029215	-0.222208	-0.007739	-0.017288	0.005383	1.000000	0.566006	0.117695	0.102203	0.088377
diff_created_lastorder	-0.001032	0.212278	0.021469	0.015295	0.207144	0.035061	-0.284345	-0.007707	0.010014	0.004021	0.566006	1.000000	0.885287	0.236647	0.589402
diff_firstorder_lastorder	0.009165	0.217064	0.035987	0.018232	0.200393	0.025753	-0.217167	-0.004918	0.021815	0.001807	0.117695	0.885287	1.000000	0.227405	0.660124
log_avgorder	0.019206	0.147774	-0.015881	-0.025602	0.809817	0.072540	-0.149069	-0.049000	0.053648	0.002455	0.102203	0.236647	0.227405	1.000000	0.259689
log_ordfreq	0.019391	0.221813	0.023881	0.026001	0.212719	0.180555	-0.265753	-0.014872	0.035646	-0.003347	0.088377	0.589402	0.660124	0.259689	1.000000

Exploratory Data Analysis
-Regression on 'retained'

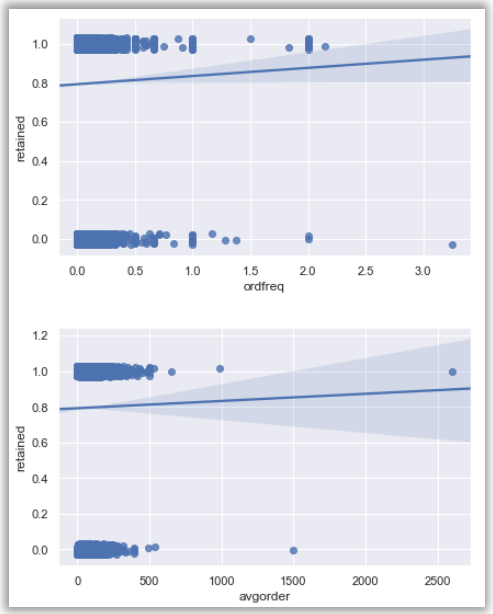
Emails:
- Esent: Strong positive relationship
- Others, low to moderate relationship



Dates:
- Days from **created to first order** shows a moderate **negative** relationship

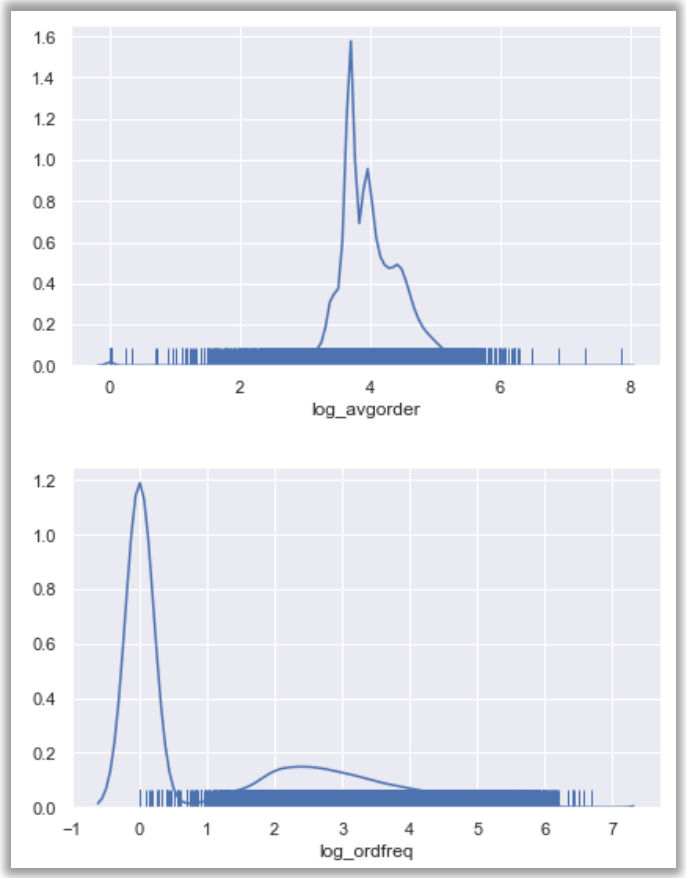


Orders:
- Outliers distort the relationship, however orderfreq shows a moderate positive relationship



Exploratory Data Analysis
-Distributions

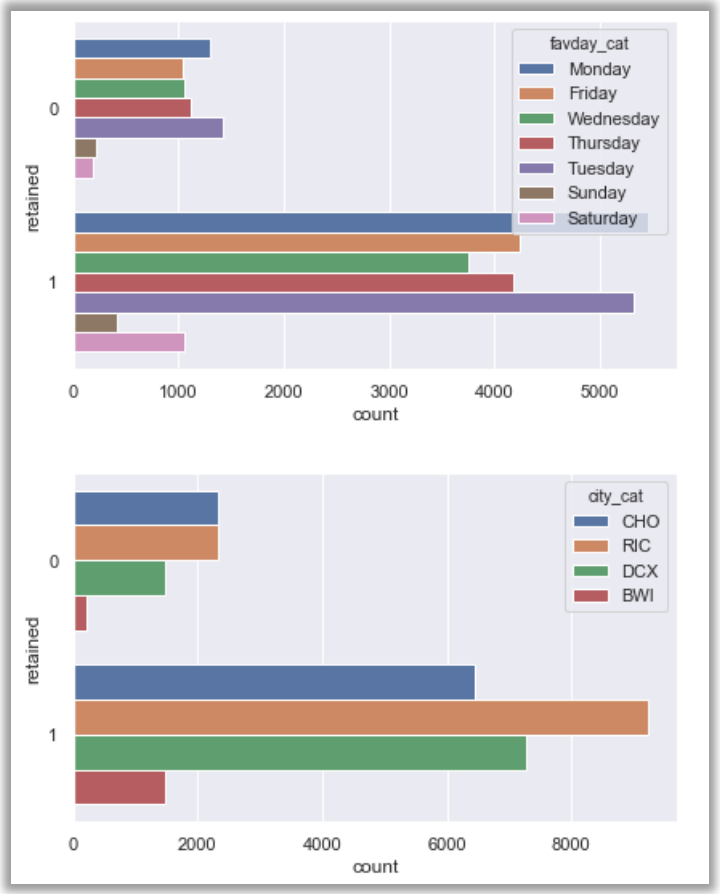
- Average orders is a normal distribution, with outliers



- Order frequency is not a normal distribution, but a left tail

- **Tuesday's** appear to be the **most** popular **favorite** purchase day
- **Sunday's** are the **least favorite** purchase day
  - Possibly due to e-mail campaign strategy
  - Begin on **Sundays**

- **Richmond** shows the **highest retention** rate. This is most likely due to the '**Richmond Expansion**' and birthplace of **Charlottesville**.



Models – Types & Parameters
DecisionTree | RandomForest

Decision Tree

- Sklearn.tree.**DecisionTreeClassifier**
- **Cross Fold Validation** = 3
- **scoring** = [‘precision_macro’, ‘recall_macro’]
- **max_depth** = None
- **min_samples_split** = 2
- **criterion** = ‘gini’ (measure the quality of split – Gini impurity)
- **splitter** = ‘best’
- **max_features** = None

Random Forest

- Sklearn.ensemble.**RandomForestClassifier**
- **Cross Fold Validation** = 3
- **scoring** = [‘precision_macro’, ‘recall_macro’]
- **n_estimators = 100**
- **Criterion** = "gini“
- **max_depth**= None
- **min_samples_split** = 2
- **min_samples_leaf** = 1
- **min_weight_fraction_leaf**=0.0
- **max_features**="auto“

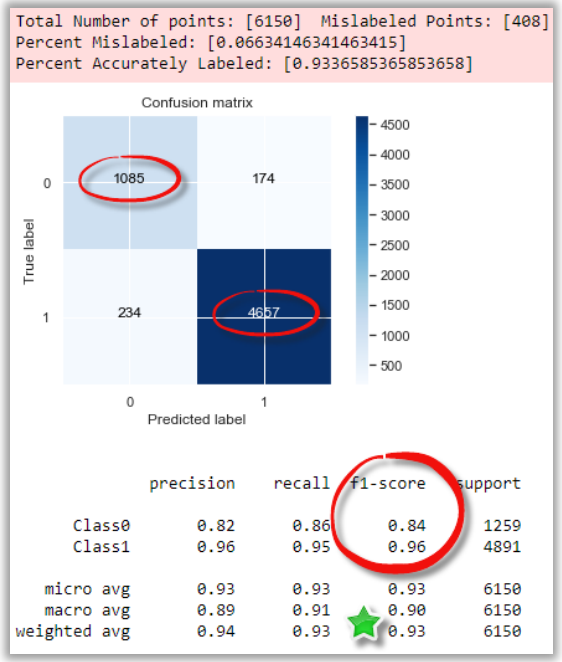
Model Features: **'esent', 'eopenrate', 'eclickrate', 'paperless', 'refill', 'doorstep', 'diff_created_firstorder', 'diff_created_lastorder', 'diff_firstorder_lastorder', 'log_avgorder', 'log_ordfreq'**

Models - Results

DecisionTree | RandomForest

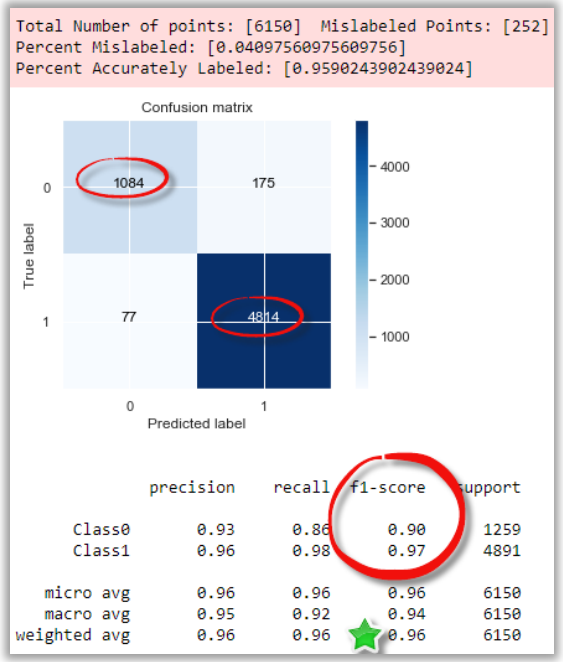
Decision Tree

- Train Hit Rate: 99%
- Test Hit Rate: 93%

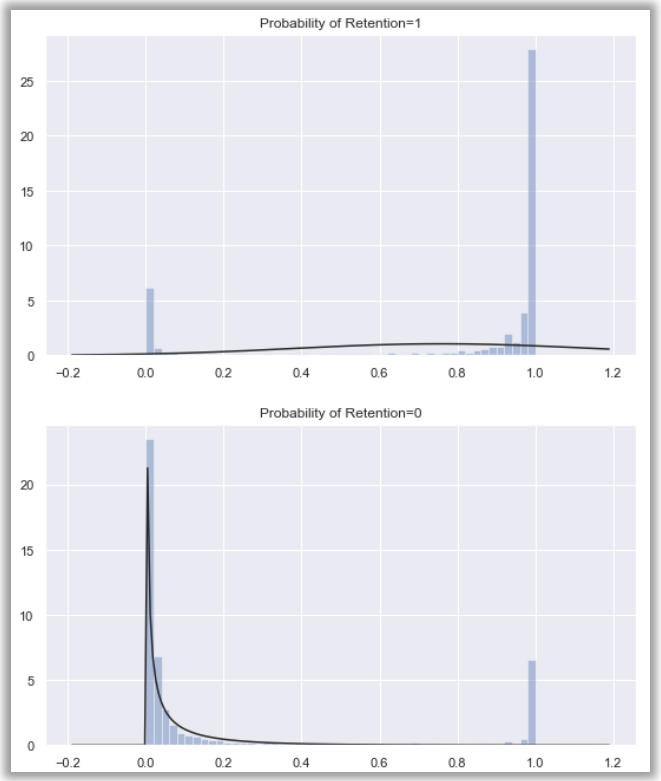


Random Forest

- Train Hit Rate: 99.6%
- Test Hit Rate: 96%

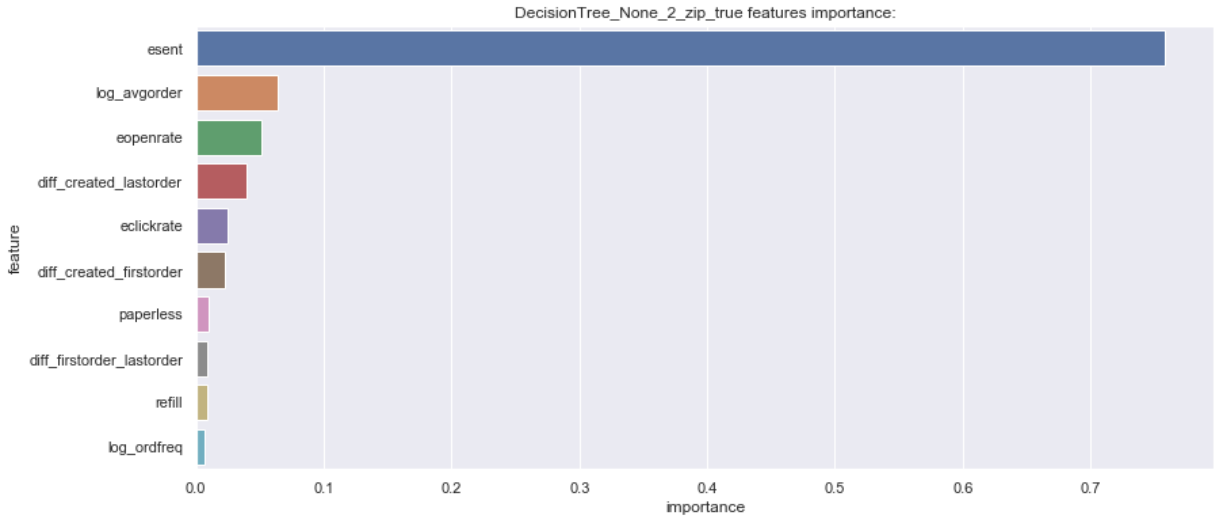


Predicted Probability Distributions

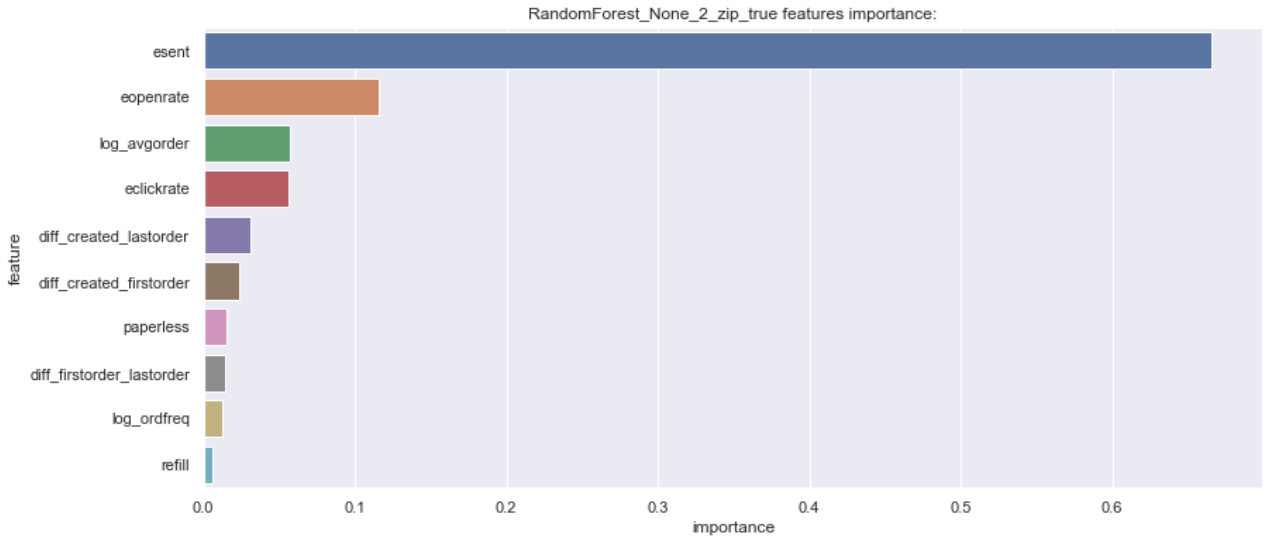


Models – Feature Importance
DecisionTree | RandomForest

Decision Tree



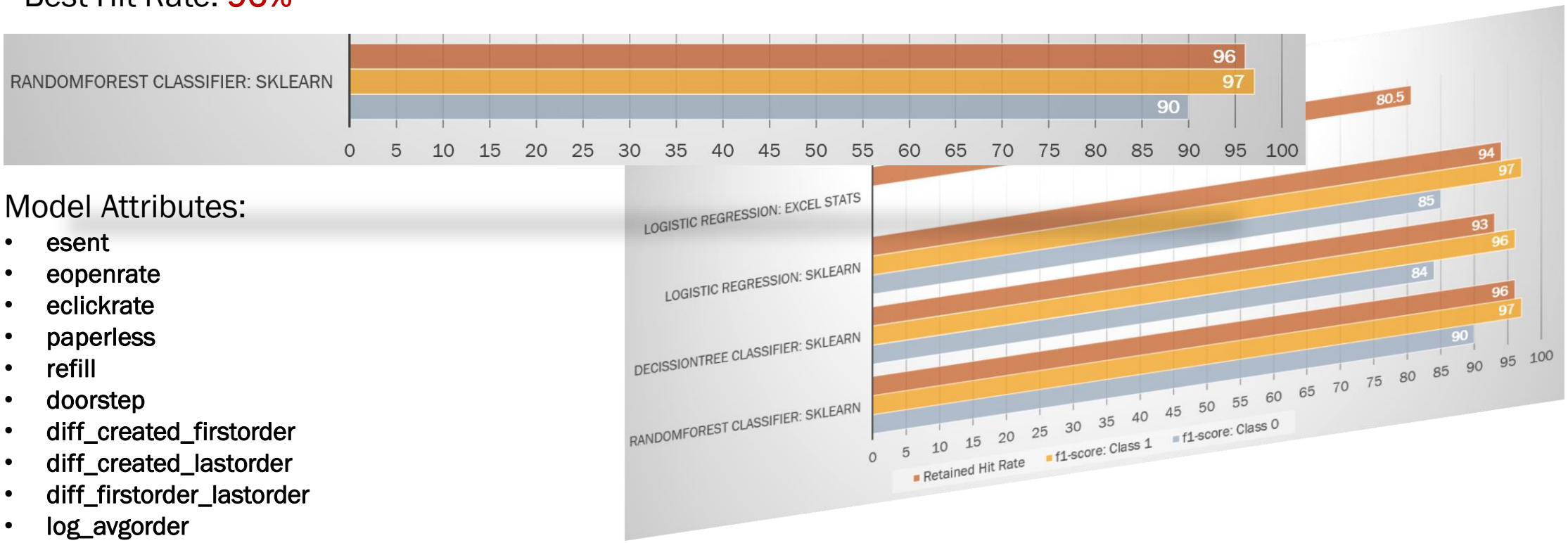
Random Forest



- Esent in all models and experiments had the **greatest impact** to **predicting 'retained'**
  - This is to be expected based on the **promotional** investments Retail Relay made with their **e-mail and social media campaigns**

Conclusion

Best Hit Rate: 96%



Model Attributes:

- esent
- eopenrate
- eclickrate
- paperless
- refill
- doorstep
- diff_created_firstorder
- diff_created_lastorder
- diff_firstorder_lastorder
- log_avgorder
- log_ordfreq

Recommendations

The e-mail marketing campaigns proved to be significantly successful, above all else, continued campaigns using the same medium should be explored along with other social media and promotional option exploration.

Retail Relay

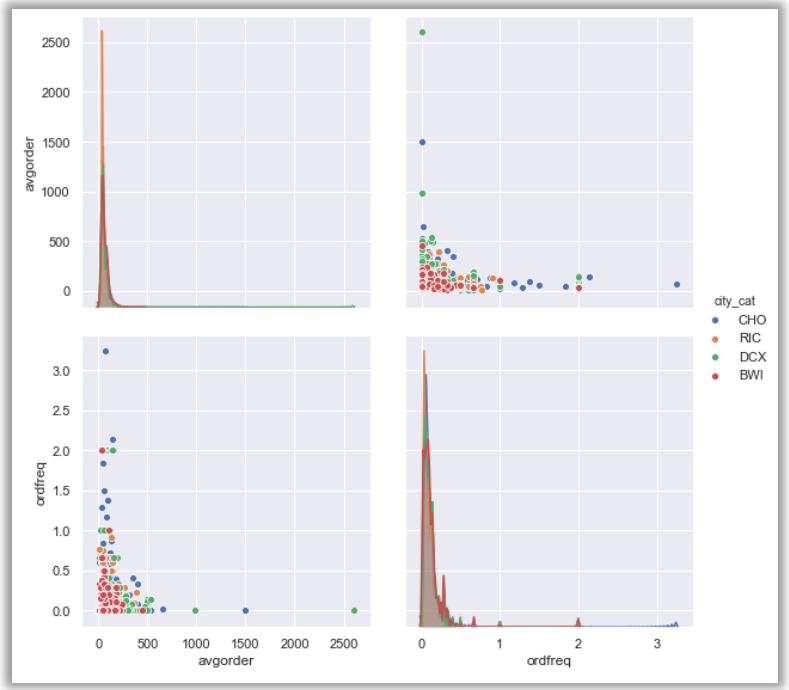
Defection Detection:

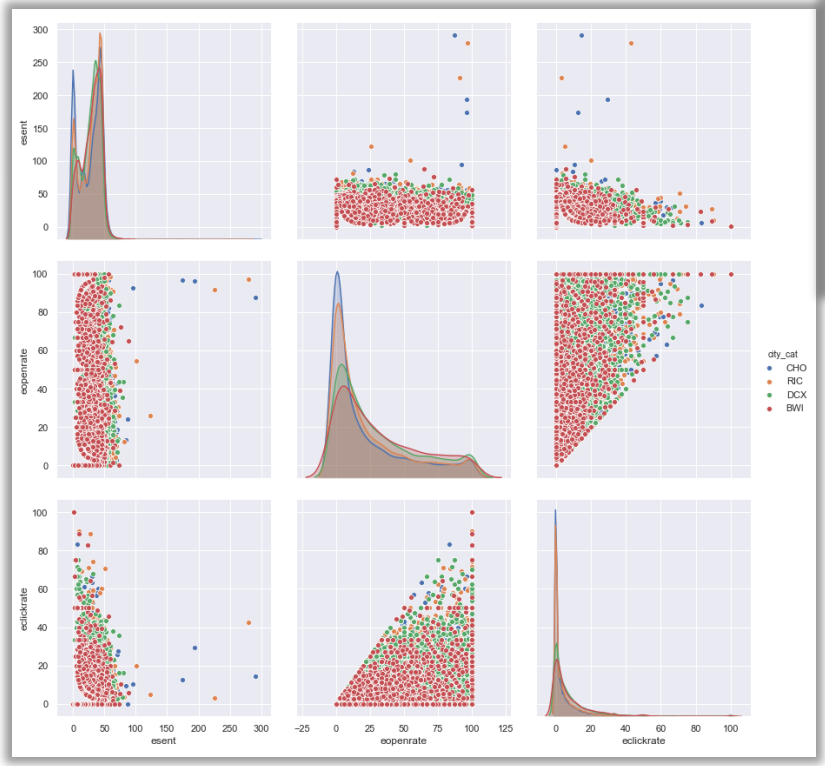
Measuring and Understanding the Predictive Accuracy of Customer Churn Models

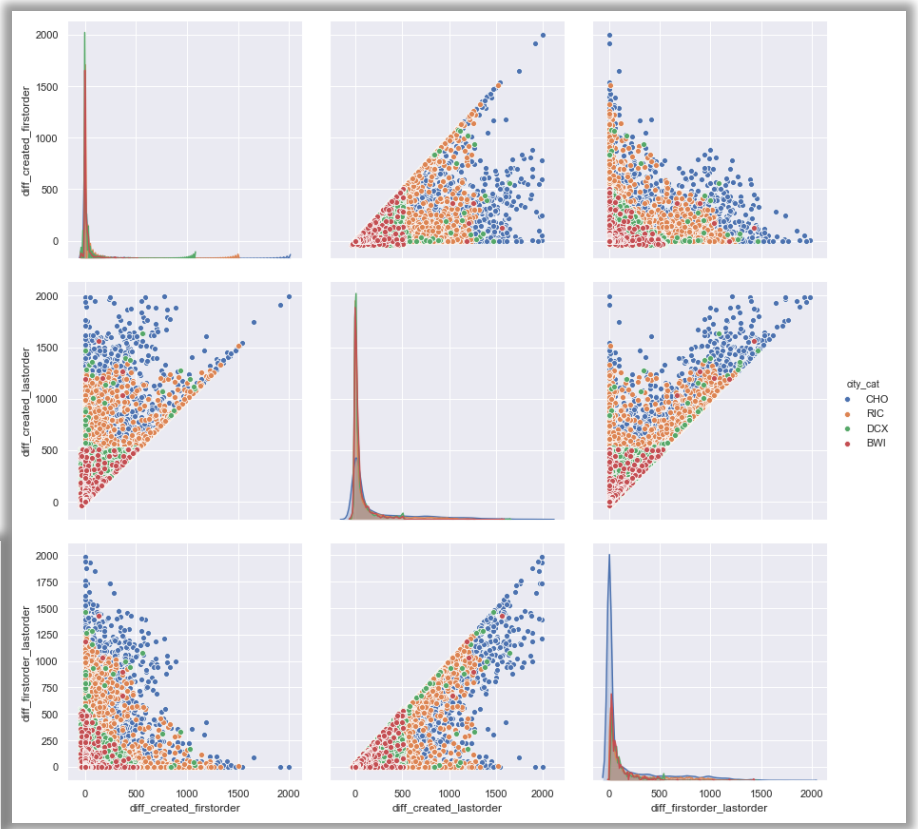
Retail Relay © case study

Assignment Questions:

1. Use the **Relay train data** to develop a model to predict customer retention.
    1. Use '**logistic regression**' to **predict** the variable '**retained**.'
    2. Use any combination of the independent variables available in the data to obtain a model with the best predictive ability and usability.
2. Using the best fit model, **predict retention** in the **test data**.
    1. Use the **coefficients** obtained from the model estimated using the train data to do this. Name this predicted value '**pretrain**.'
3. **Calculate** the **hit rate**. This can be calculated as **% of matches** between the value of '**pretain**' and '**retained**' in the train data.
4. Present results in class.


avgorder
ordfreq
city_cat
CHO
RIC
DCX
BWI


esent
eopenrate
eclickrate
city_cat
CHO
RIC
DCX
BWI


diff_created_firstorder
diff_created_lastorder
diff_firstorder_lastorder
city_cat
CHO
RIC
DCX
BWI

OLS Regression Results

Dep. Variable:	retained	R-squared:	0.576
Model:	OLS	Adj. R-squared:	0.575
Method:	Least Squares	F-statistic:	1399.
Date:	Mon, 24 Feb 2020	Prob (F-statistic):	0.00
Time:	10:35:40	Log-Likelihood:	-1746.9
No. Observations:	20643	AIC:	3536.
Df Residuals:	20622	BIC:	3702.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2610	0.012	21.762	0.000	0.237	0.285
favday_cat[T.Monday]	-0.0198	0.006	-3.347	0.001	-0.031	-0.008
favday_cat[T.Saturday]	-0.0256	0.011	-2.371	0.018	-0.047	-0.004
favday_cat[T.Sunday]	0.0006	0.015	0.044	0.965	-0.028	0.029
favday_cat[T.Thursday]	-0.0183	0.006	-2.901	0.004	-0.031	-0.006
favday_cat[T.Tuesday]	-0.0234	0.006	-3.929	0.000	-0.035	-0.012
favday_cat[T.Wednesday]	-0.0136	0.006	-2.112	0.035	-0.026	-0.001
city_cat[T.CHO]	-0.0119	0.009	-1.268	0.205	-0.030	0.007
city_cat[T.DCX]	-0.0213	0.009	-2.410	0.016	-0.039	-0.004
city_cat[T.RIC]	-0.0313	0.009	-3.434	0.001	-0.049	-0.013
esent	0.0184	0.000	160.109	0.000	0.018	0.019
eopenrate	0.0017	7.58e-05	21.932	0.000	0.002	0.002
eclickrate	0.0011	0.000	5.086	0.000	0.001	0.001
avgorder	-0.0003	4.51e-05	-7.214	0.000	-0.000	-0.000
ordfreq	-0.0768	0.018	-4.307	0.000	-0.112	-0.042
paperless	0.0743	0.004	16.763	0.000	0.066	0.083
refill	0.0399	0.007	5.994	0.000	0.027	0.053
doorstep	0.0117	0.010	1.151	0.250	-0.008	0.032
train	0.0031	0.005	0.672	0.501	-0.006	0.012
diff_created_firstorder	-0.0001	1.51e-05	-6.639	0.000	-0.000	-7.05e-05
diff_firstorder_lastorder	-0.0002	9.01e-06	-27.280	0.000	-0.000	-0.000

Omnibus:	6120.553	Durbin-Watson:	1.520
Prob(Omnibus):	0.000	Jarque-Bera (JB):	94875.062
Skew:	-1.006	Prob(JB):	0.00
Kurtosis:	13.308	Cond. No.	2.65e+03

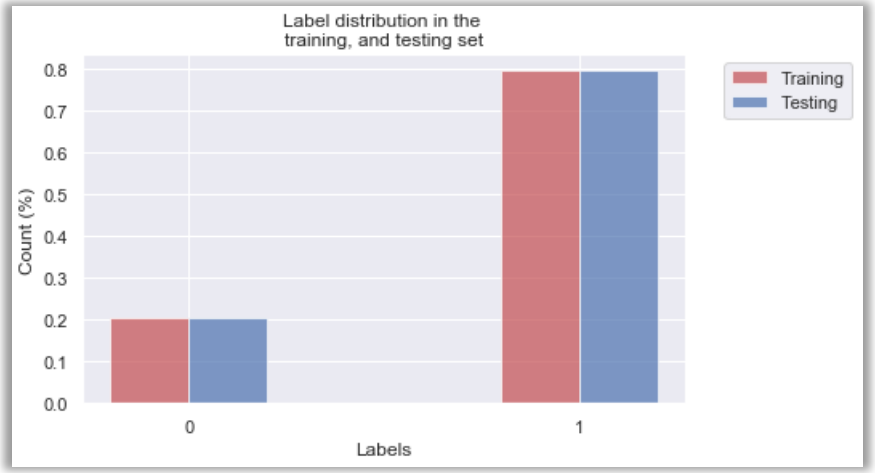
OLS Regression Results

Dep. Variable:	retained	R-squared:	0.573
Model:	OLS	Adj. R-squared:	0.573
Method:	Least Squares	F-statistic:	3962.
Date:	Mon, 24 Feb 2020	Prob (F-statistic):	0.00
Time:	10:35:41	Log-Likelihood:	-1803.5
No. Observations:	20643	AIC:	3623.
Df Residuals:	20635	BIC:	3686.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2225	0.005	41.674	0.000	0.212	0.233
esent	0.0183	0.000	160.498	0.000	0.018	0.018
eopenrate	0.0018	6.49e-05	28.488	0.000	0.002	0.002
avgorder	-0.0003	4.5e-05	-7.311	0.000	-0.000	-0.000
ordfreq	-0.0708	0.018	-3.983	0.000	-0.106	-0.036
paperless	0.0813	0.004	19.503	0.000	0.073	0.089
refill	0.0414	0.006	6.394	0.000	0.029	0.054
diff_firstorder_lastorder	-0.0002	8.8e-06	-27.663	0.000	-0.000	-0.000

Omnibus:	5944.316	Durbin-Watson:	1.504
Prob(Omnibus):	0.000	Jarque-Bera (JB):	89071.985
Skew:	-0.976	Prob(JB):	0.00
Kurtosis:	12.987	Cond. No.	2.38e+03

```
y shape: (30747,)
X shape: (30747, 11)
X_train shape: (24597, 11)
X_test shape: (6150, 11)
y_train shape: (24597,)
y_test shape: (6150,)
```



Models – Parameters
Logistic Regression

Logistic Regression – sklearn LogisticRegressionCV

- Cross Fold Validation: 5
- Solver: lbfgs – algorithm used in the optimization problem
- Max_iter: 100
- Penalty: l2

Model Features: 'esent', 'eopenrate', 'eclickrate', 'paperless', 'refill', 'doorstep', 'diff_created_firstorder', 'diff_created_lastorder', 'diff_firstorder_lastorder', 'log_avgorder', 'log_ordfreq'

Models - Results
Logistic Regression

Logistic Regression – Hit Rate Prediction Accuracy

- Train Hit Rate: 97%
- Test Hit Rate: 94.4%

- Percent Accuracy – True **Class 0**: 85%
- Percent Accuracy – True **Class 1**: 97%

❖ R^2 Coefficient of determination: 0.656434010434623 3

	coef	coef_value
0	intercept	-1.822325
1	esent	0.211279
2	eopenrate	0.007011
3	eclickrate	0.010719
4	paperless	0.190017
5	refill	0.743016
6	doorstep	0.858336
7	diff_created_firstorder	-0.000297
8	diff_created_lastorder	-0.001047
9	diff_firstorder_lastorder	-0.000750
10	log_avgorder	-0.179801
11	log_ordfreq	-0.162819