Report: In-Depth Analysis

- ❏ A 2-3 page report on the steps and findings from machine learning in-depth analysis, uploaded to GitHub.
- ❏ The submission includes a justification of the machine learning technique, and features selection and evaluation metrics/techniques utilized.

Summary:

- My goal was to understand what features were important to predicting whether an intro call would be qualified.
- In order to classify whether intro calls would be classified, I built three models: a logistic regression model, a random forests model, and a gradient boosted model.
- Leveraging different feature engineering techniques and hyperparameter tuning, I was able to attain 79.3 % accuracy in classifying Intro Call qualifications (table of results shown below) with the Gradient Boosted 1-Hot Encoded Model with Hyperparameter tuning.
- The top 5 features across all three models in determining Intro Call Qualification Status included:
    - inferScore___Lead_AddedInfo
    - totalEMails___Lead_AddedInfo
    - totalCalls___Lead_AddedInfo
    - introCallCreated_leadCreated_delta
    - assignedToRole___IntroCall_OtherInfo_map

- The features I assumed would be highly ranked but weren't included:
    - country___Lead_LeadCompanyInformation_map
    - trafficChannel___Lead_MarketingInformation_map_map

- product2___IntroCall_MeetingDetails_WalkMe
- Interestingly the top performing model also departed from the other models in terms of features ranked #5-#10 (see screenshot below). .

Summary of Results:

Model	Version	Performance	Performance with Param. Tuning	Optimal Params	Top Features
Logistic Regression	Products - 1 Hot Encoded	72.6%	72.6%	{'C': 1, 'max_iter': 100}	
	Products - Not 1 Hot Encoded	67.8%			
Random Forest	Products - 1 Hot Encoded	74.4%	79.0%	{'bootstrap': False, 'max_depth': 60, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 400}	1. inferScore___Lead_AddedInfo 2. totalEMails___Lead_AddedInfo 3. totalCalls___Lead_AddedInfo 4. introCallCreated_leadCreated_delta 5. assignedToRole___IntroCall_OtherInfo_map 6. mnth_createddate___IntroCall_ImportantSystemInfo_clean 7. mnth_createddate___Lead_ImportantSystemInfo_clean

					8. country___Lead_LeadCompanyInformation_map 9. trafficChannel___Lead_MarketingInformation_map_map 10. product2___IntroCall_MeetingDetails_WalkMe
	Products - Not 1 Hot Encoded	70.7%			1. inferScore___Lead_AddedInfo 2. totalEMails___Lead_AddedInfo 3. totalCalls___Lead_AddedInfo 4. introCallCreatedleadCreated_delta 5. assignedToRole___IntroCall_OtherInfo_map 6. mnth_createddate___IntroCall_ImportantSystemInf 7. mnth_createddate___Lead_ImportantSystemInfo_clean 8. year_createddate___IntroCall_ImportantSystemInf 9. trafficChannel___Lead_MarketingInformation_map_map 10. country___Lead_LeadCompanyInformation_map
Gradient Boosted	Products - 1 Hot Encoded	76.7%	79.3%	{'colsample_bytree': 1.0, 'gamma': 1, 'max_depth': 6, 'min_child_weight': 9, 'subsample': 1.0}	• year_createddate___IntroCall_ImportantSystemInf • totalEMails___Lead_AddedInfo • inferScore___Lead_AddedInfo • introCallCreatedleadCreated_delta

				• product2___IntroCall_MeetingDetails_WalkMe • totalCalls___Lead_AddedInfo • customerType___Lead_LeadCompanyInformation_map • mnth_createddate___IntroCall_ImportantSystemInf • year_createddate___Lead_ImportantSystemInfo_clean • customerOrEmployee___IntroCall_MeetingDetails_map
	Products - Not 1 Hot Encoded	73.5%		• totalEMails___Lead_AddedInfo • introCallCreated_leadCreated_delta • totalCalls___Lead_AddedInfo • year_createddate___IntroCall_ImportantSystemInf… • inferScore___Lead_AddedInfo • customerType___Lead_LeadCompanyInformation_map • assignedToRole___IntroCall_OtherInfo_map • customerOrEmployee___IntroCall_MeetingDetails_map • decisionMaker___IntroCall_MeetingDetails_map • mnth_createddate___Lead_ImportantSystemInfo_clean

Top 10 Features by model and rank:

	Random Forest - 1 Hot	Random Forest - Not Hot	Gradient Boosted - 1 Hot	Gradient Boosted - Not Hot
inferScore___Lead_AddedInfo	1	1	3	5
totalEMails___Lead_AddedInfo	2	2	2	1
totalCalls___Lead_AddedInfo	3	3	6	3
introCallCreated_leadCreated_delta	4	4	4	2
assignedToRole___IntroCall_OtherInfo_map	5	5		7
mnth_createddate___IntroCall_ImportantSystemInfo_clean	6			
mnth_createddate___Lead_ImportantSystemInfo_clean	7	7		10
country___Lead_LeadCompanyInformation_map	8	10		
trafficChannel___Lead_MarketingInformation_map_map	9	9		
product2___IntroCall_MeetingDetails_WalkMe	10		5	
mnth_createddate___IntroCall_ImportantSystemInf		6	8	
year_createddate___IntroCall_ImportantSystemInf		8	1	
customerType___Lead_LeadCompanyInformation_map			7	6
year_createddate___Lead_ImportantSystemInfo_clean			9	
customerOrEmployee___IntroCall_MeetingDetails_map			10	8
year_createddate___IntroCall_ImportantSystemInf...				4
decisionMaker___IntroCall_MeetingDetails_map				9

Data Preparation for Model Building:

- Each model used the same starting dataset (), with all categorical values numerically encoded by creating dictionaries of corresponding values and mapping key-value pairs of through a custom function (**clean_map(df, dictToMap, oldColName)**).

- I produced two datasets to compare the difference in performance for the models when data (like the products list) was 1-Hot Encoded versus cleaned and separated out into unique product identifiers. This change in feature engineering would have impacted prospects with multiple products listed (which would likely have been Strategic or Enterprise accounts), along with Mid-Market accounts that included support services.
- **closedIntroCalls_Data_logisticRegression** was used to test the performance of each model when the products column was 1-Hot Encoded, **closedIntroCalls_wProducts_Data_logisticRegression_products** was used to test the performance when products are individually broken out.

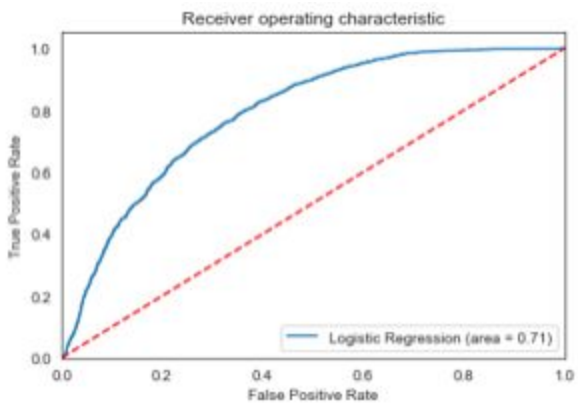
Approach to Model Building & Tuning:

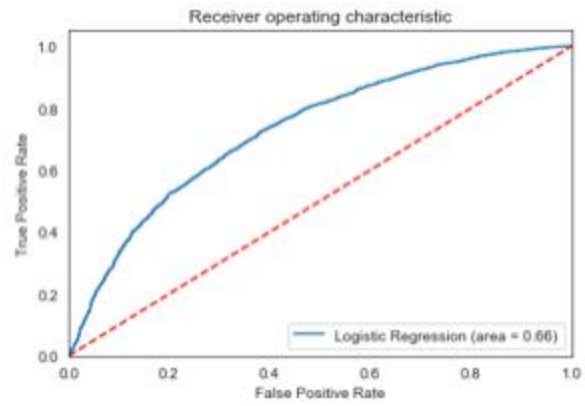
- For each model I first ran each version of the starting data sets, creating a test and train split via **sklearn's trest_train_split function**. The test and train sets were scaled with **StandardScaler** and a new classifier instantiated (**LogisticRegression(), RandomForestClassifier(), XGBClassifier()**). The models were then trained using the training sets and the classifier objects **.fit** method, after which they were then used to predict the test data labels and evaluated for performance.
- For all the models, the 1 Hot encoded version of the data performed the best. In order to further improve the model performance I then used **RandomizedSearchCV** to search for the best hyperparameters and then used **GridSearchCV** to further refine the selection of parameters. Especially for the RandomForestClassifier, using RandomizedSearchCV to sample possible combinations of parameters to narrow down the search for the optimal parameters was essential, given all the possible parameters available.

Model 1: Logistic Regressions - Detailed Analysis

Bazeley, Mikiko
Springboard 2018

Capstone Project 1: Scoring First Sales Demo's





Logistic Regression Model
(Version 1) - Products are 1 Hot Encoded

Accuracy Score: 0.726498696785404

Logistic Regression Model
(Version 2) - Products not Hot Encoded

```
Accuracy Score:  0.6783666377063423
```

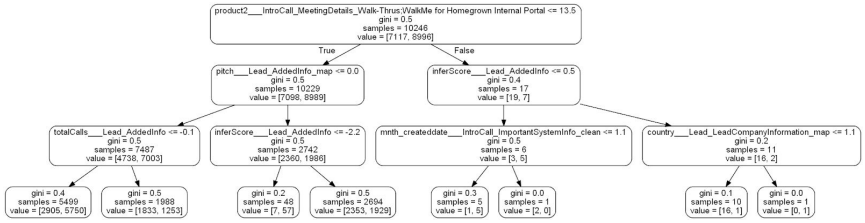
```
Params:  <bound method BaseEstimator.get_params of
LogisticRegression(C=1.0, class_weight=None,
dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100,
multi_class='warn',
          n_jobs=None, penalty='l2',
random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)>
```

```
Params:  <bound method BaseEstimator.get_params of
LogisticRegressionCV(Cs=10, class_weight=None,
cv='warn', dual=False,
           fit_intercept=True,
intercept_scaling=1.0, max_iter=100,
           multi_class='warn', n_jobs=None,
penalty='l2',
           random_state=None, refit=True,
scoring=None, solver='lbfgs',
           tol=0.0001, verbose=0)>
```

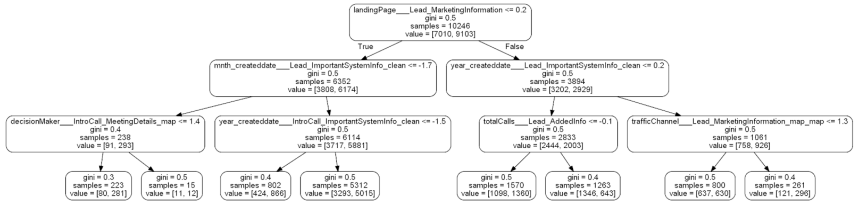
Best parameter: {'C': 1, 'max_iter': 100}
Best score: 0.7299003707136237
Test set accuracy: 0.726498696785404

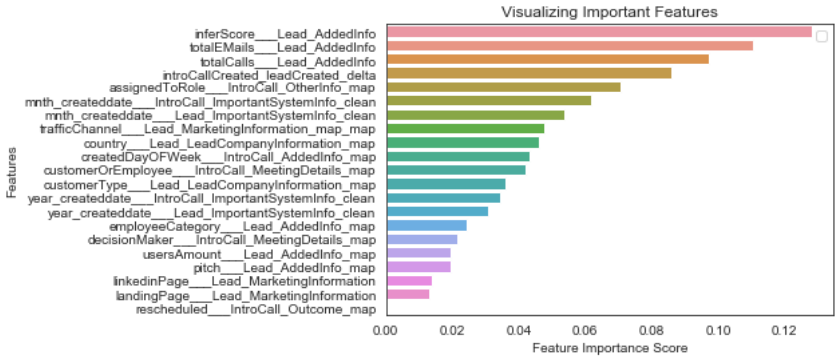
Model 2: Random Forest- Detailed Analysis

Version 1 Hot Encoded



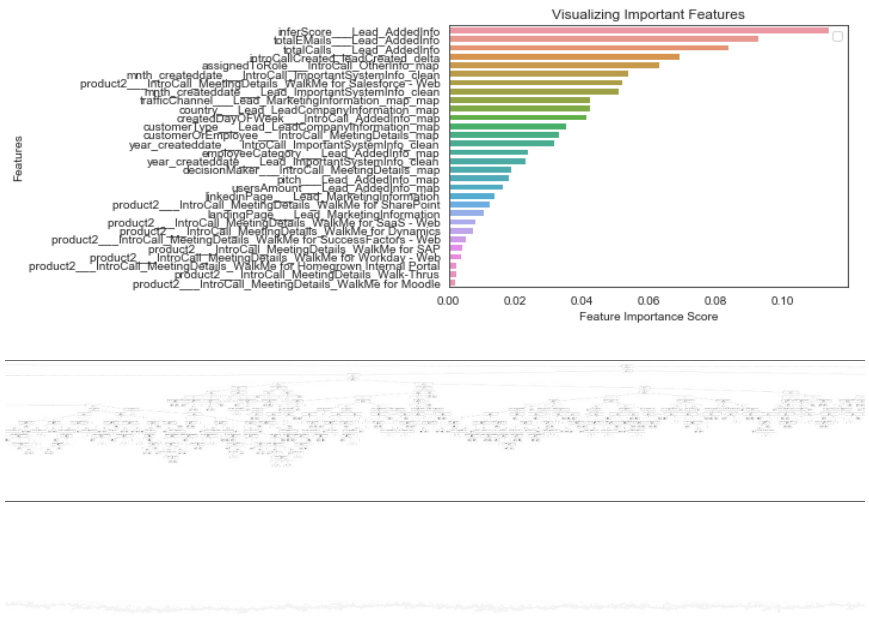
Version 2 Hot Encoded

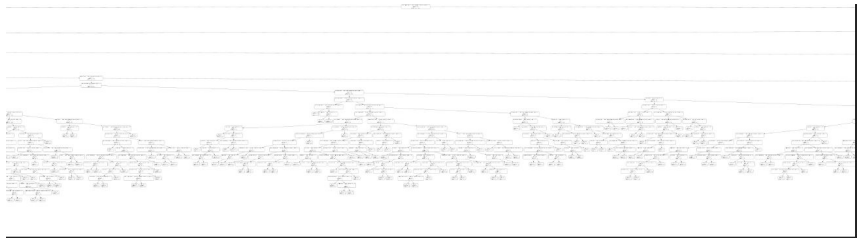




Bazeley, Mikiko
Springboard 2018

Capstone Project 1: Scoring First Sales Demo's





Random Forest Classifier (Version 1) - Products are 1 Hot Encoded

```
Accuracy Score:  0.7437011294526499


Confusion Matrix (Test Set):
[[2224  830]
 [ 940 2912]]


Confusion Matrix (Train Set):
[[7009   71]
 [  89 8944]]
```

Random Forest Classifier (Version 2) - Products not Hot Encoded

```
Accuracy Score:  0.7066319142774399


Confusion Matrix (Test Set):
[[2084  929]
 [1097 2796]]

Confusion Matrix (Train Set):
[[7050   71]
 [ 147 8845]]
```

Params being used:

```
{'bootstrap': True,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 10,
 'n_jobs': None,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False
```

Params Currently in Use::

```
{'bootstrap': True,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 10,
 'n_jobs': None,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```

GridSearchCV:

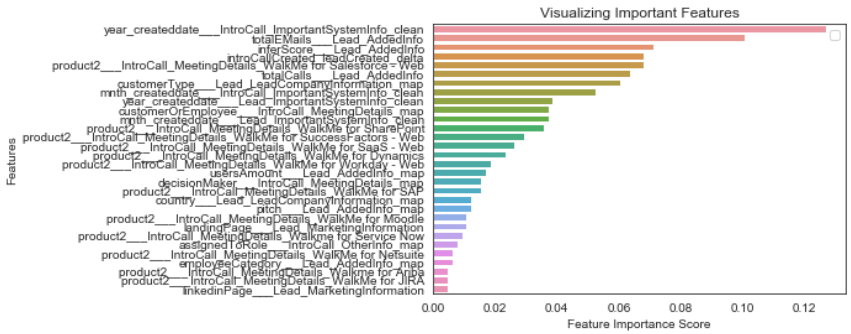
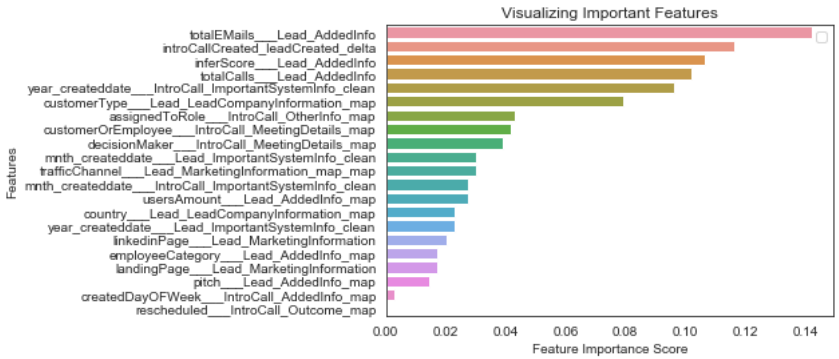
```
GridSearchCV(cv=3,
error_score='raise-deprecating',

estimator=RandomForestClassifier(bootstrap=True,
class_weight=None, criterion='gini',
            max_depth=None, max_features='auto',
max_leaf_nodes=None,
            min_impurity_decrease=0.0,
min_impurity_split=None,
            min_samples_leaf=1,
min_samples_split=2,
```

min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None, oob_score=False, random_state=None, verbose=0, warm_start=False), fit_params=None, iid='warn', n_jobs=-1, param_grid={'n_estimators': [400], 'min_samples_split': [7, 10, 13], 'min_samples_leaf': [1, 3, 5], 'max_features': ['sqrt'], 'max_depth': [20, 40, 50, 60], 'bootstrap': [False]}, pre_dispatch='2*n_jobs', refit=True, return_train_score='warn', scoring=None, verbose=2)	
Best parameter: {'bootstrap': False, 'max_depth': 60, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 400} Best score: 0.7901694284118413 Test set accuracy: 0.7903272516652187	
Feature Imp: inferScore___Lead_AddedInfo 0.106930573 totalEMails___Lead_AddedInfo 0.095770247 totalCalls___Lead_AddedInfo 0.089220348 introCallCreated_leadCreated_delta 0.068788421 assignedToRole___IntroCall_OtherInfo_map 0.06388306 mnth_createddate___IntroCall_ImportantSystemInfo_clean	Feature Imp: inferScore___Lead_AddedInfo 0.125791 totalEMails___Lead_AddedInfo 0.113275 totalCalls___Lead_AddedInfo 0.098810

0.058059835 mnth_createddate___Lead_ImportantSystemInfo_clean 0.050737931 country___Lead_LeadCompanyInformation_map 0.046237732 trafficChannel___Lead_MarketingInformation_map_map 0.046013478 product2___IntroCall_MeetingDetails_WalkMe for Salesforce - Web 0.04367586 createdDayOFWeek___IntroCall_AddedInfo_map 0.041720427 customerType___Lead_LeadCompanyInformation_map 0.036744751 customerOrEmployee___IntroCall_MeetingDetails_map 0.029965182 year_createddate___IntroCall_ImportantSystemInfo_clean 0.02616131 year_createddate___Lead_ImportantSystemInfo_clean 0.024584646 employeeCategory___Lead_AddedInfo_map 0.023754743 pitch___Lead_AddedInfo_map 0.018499198 decisionMaker___IntroCall_MeetingDetails_map 0.018047798 usersAmount___Lead_AddedInfo_map 0.016498403	introCallCreated_leadCreated_delta 0.087681 assignedToRole___IntroCall_OtherInfo_map 0.072087 mnth_createddate___IntroCall_ImportantSystemInf... 0.059484 mnth_createddate___Lead_ImportantSystemInfo_clean 0.053333 year_createddate___IntroCall_ImportantSystemInf... 0.045717 trafficChannel___Lead_MarketingInformation_map_map 0.044573 country___Lead_LeadCompanyInformation_map 0.044475 createdDayOFWeek___IntroCall_AddedInfo_map 0.043441 customerType___Lead_LeadCompanyInformation_map 0.037227 customerOrEmployee___IntroCall_MeetingDetails_map 0.037219 year_createddate___Lead_ImportantSystemInfo_clean 0.027207 employeeCategory___Lead_AddedInfo_map 0.025754 decisionMaker___IntroCall_MeetingDetails_map 0.020911 usersAmount___Lead_AddedInfo_map 0.019178 pitch___Lead_AddedInfo_map 0.016852 landingPage___Lead_MarketingInformation 0.014002 linkedinPage___Lead_MarketingInformation 0.012983 rescheduled___IntroCall_Outcome_map

	0.000000

Model 3: Gradient Boosted- Detailed Analysis

Version 1 Hot Encoded	Version 2 Hot Encoded
	
Gradient Boosted (Version 1) - Products are 1 Hot Encoded	Gradient Boosted (Version 2) - Products not Hot Encoded

```
Accuracy Score:  0.766724587315378


Confusion Matrix (Test Set):
[[1861 1144]
 [ 467 3434]]


Confusion Matrix (Train Set):
[[4637 2492]
 [ 942 8042]]
```

```
Accuracy Score:  0.7347234289024037


Confusion Matrix (Test Set):
[[1903 1155]
 [ 677 3171]]

Confusion Matrix (Train Set):
[[4576 2500]
 [1455 7582]]
```

Params Currently in Use:

```
{'base_score': 0.5,
 'booster': 'gbtree',
 'colsample_bylevel': 1,
 'colsample_bytree': 1,
 'gamma': 0,
 'learning_rate': 0.1,
 'max_delta_step': 0,
 'max_depth': 3,
 'min_child_weight': 1,
 'missing': None,
 'n_estimators': 100,
 'n_jobs': 1,
 'nthread': None,
 'objective': 'binary:logistic',
```

Params Currently in Use::

```
{'base_score': 0.5,
 'booster': 'gbtree',
 'colsample_bylevel': 1,
 'colsample_bytree': 1,
 'gamma': 0,
 'learning_rate': 0.1,
 'max_delta_step': 0,
 'max_depth': 3,
 'min_child_weight': 1,
 'missing': None,
 'n_estimators': 100,
 'n_jobs': 1,
 'nthread': None,
 'objective': 'binary:logistic',
 'random_state': 0,
 'reg_alpha': 0,
 'reg_lambda': 1,
```

'random_state': 0, 'reg_alpha': 0, 'reg_lambda': 1, 'scale_pos_weight': 1, 'seed': None, 'silent': True, 'subsample': 1}	'scale_pos_weight': 1, 'seed': None, 'silent': True, 'subsample': 1}
Best Params: Best parameter: {'colsample_bytree': 1.0, 'gamma': 1, 'max_depth': 6, 'min_child_weight': 9, 'subsample': 1.0} Best score: 0.7926518959846087 Test set accuracy: 0.7930784824790038	
Feature Imp: year_createddate___IntroCall_ImportantSystemInf... 0.126935 totalEMails___Lead_AddedInfo 0.100619 inferScore___Lead_AddedInfo 0.071207 introCallCreated_leadCreated_delta 0.068111 product2___IntroCall_MeetingDetails_WalkMe for ... 0.068111 totalCalls___Lead_AddedInfo 0.063467 customerType___Lead_LeadCompanyInformation_map	Feature Imp: 0 totalEMails___Lead_AddedInfo 0.151429 introCallCreated_leadCreated_delta 0.122857 totalCalls___Lead_AddedInfo 0.104286 year_createddate___IntroCall_ImportantSystemInf... 0.102857 inferScore___Lead_AddedInfo 0.100000 customerType___Lead_LeadCompanyInformation_map 0.072857 assignedToRole___IntroCall_OtherInfo_map

```
0.060372
mnth_createddate___IntroCall_ImportantSystemInf...
0.052632
year_createddate___Lead_ImportantSystemInfo_clean
0.038700
customerOrEmployee___IntroCall_MeetingDetails_map
0.037152
mnth_createddate___Lead_ImportantSystemInfo_clean
0.037152
product2___IntroCall_MeetingDetails_WalkMe for ...
0.035604
product2___IntroCall_MeetingDetails_WalkMe for ...
0.029412
product2___IntroCall_MeetingDetails_WalkMe for ...
0.026316
product2___IntroCall_MeetingDetails_WalkMe for ...
0.023220
product2___IntroCall_MeetingDetails_WalkMe for ...
0.018576
usersAmount___Lead_AddedInfo_map
0.017028
decisionMaker___IntroCall_MeetingDetails_map
0.015480
product2___IntroCall_MeetingDetails_WalkMe for SAP
0.015480
country___Lead_LeadCompanyInformation_map
0.012384
pitch___Lead_AddedInfo_map
0.012384
product2___IntroCall_MeetingDetails_WalkMe for ...
0.010836
landingPage___Lead_MarketingInformation
0.010836
product2___IntroCall_MeetingDetails_Walkme for ...
0.009288
```

```
0.055714
customerOrEmployee___IntroCall_MeetingDetails_map
0.038571
decisionMaker___IntroCall_MeetingDetails_map
0.037143
mnth_createddate___Lead_ImportantSystemInfo_clean
0.032857
year_createddate___Lead_ImportantSystemInfo_clean
0.030000
mnth_createddate___IntroCall_ImportantSystemInf...
0.027143
landingPage___Lead_MarketingInformation
0.021429
usersAmount___Lead_AddedInfo_map
0.021429
pitch___Lead_AddedInfo_map
0.018571
trafficChannel___Lead_MarketingInformation_map_map
0.017143
linkedinPage___Lead_MarketingInformation
0.017143
country___Lead_LeadCompanyInformation_map
0.012857
employeeCategory___Lead_AddedInfo_map
0.010000
createdDayOFWeek___IntroCall_AddedInfo_map
0.005714
rescheduled___IntroCall_Outcome_map
0.000000
```

assignedToRole___IntroCall_OtherInfo_map 0.007740 product2___IntroCall_MeetingDetails_WalkMe for ... 0.006192 employeeCategory___Lead_AddedInfo_map 0.006192 product2___IntroCall_MeetingDetails_Walkme for ... 0.004644 product2___IntroCall_MeetingDetails_WalkMe for ... 0.004644 linkedinPage___Lead_MarketingInformation 0.004644	