PHONE TEAM SALES LEAD SCORING

Expanding prediction with online attribution and phone call data

Goal: Rank sales leads for the phone team to prioritize calls

Hypothesis: Demographic, online activity, and phone data are predictive of customer signups

Assumption: Contacting people most likely to sign up provides the best return on phone team effort

Current data sets

Policies
(42946 x 15)
account_number
product_state
policy_date_entered
current_td_program_name
has_tnc
policy_feature_group
product_enum
credit_score
prior_insurance
prior_bi_limit
prior_insurance_company
prior_insurance_premium
prior_ins_length_of_time
prior_liability_c
quote_status

Converted
(7791x 2)
account_number
esign_datetime

Vehicles
(51534 x10)
account_number
vehicle_id_c
make
model
year
ownership_type
loan_lending_company
current_total_daily_base
current_total_per_mile
reported_prior_yearly_mileage

Drivers
(61049 x 12)
account_number
driver_id
primary_address_postalcode
birthdate
marital_status
sex
driver_type
education_code
age_licensed
occupation_code
residence_status
total_points

Additional data sets

Attribution
(386877 x 6)
account_number
mm_category
source
medium
campaign
weblog_ts

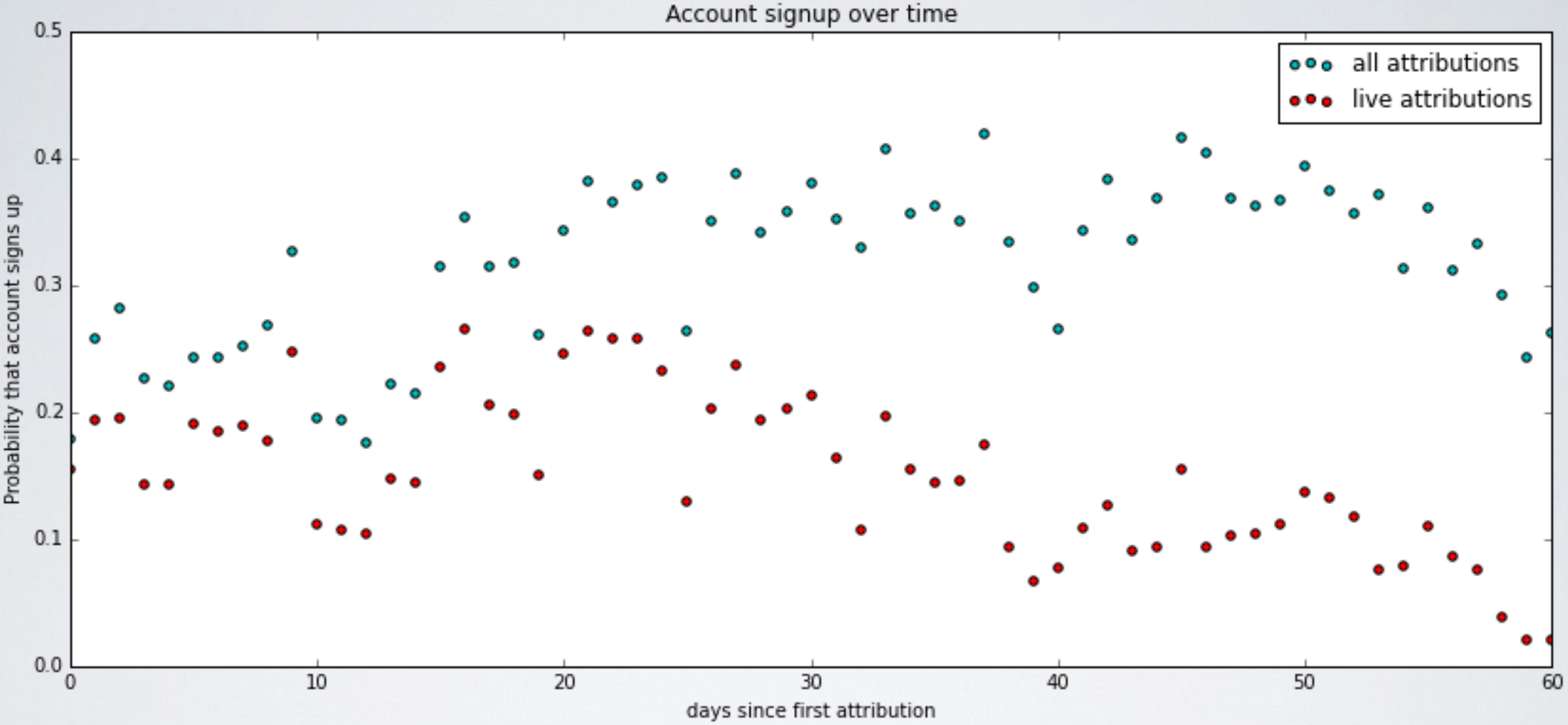
Phone
(26848 x 3)
call_time
call_result
account_number

External Zipcode Metadata
(81831 x 9)
Zipcode
State
EstimatedPopulation
TaxReturnsFiled
TotalWages
2010 Population
Land-Sq-Mi
Density Per Sq Mile
Unemp. Rate

DATA CLEANING

- Data oriented around unique accounts. (e.g. Primary driver, first policy)
- Group small categories as "other"
- Label encoding categorical features with ordinal relationship
- Dummy / one-hot encoding for categorical features
- Converted zip code to related numerical representations (population density, average income)

LEAKAGE: ONLINE ATTRIBUTION

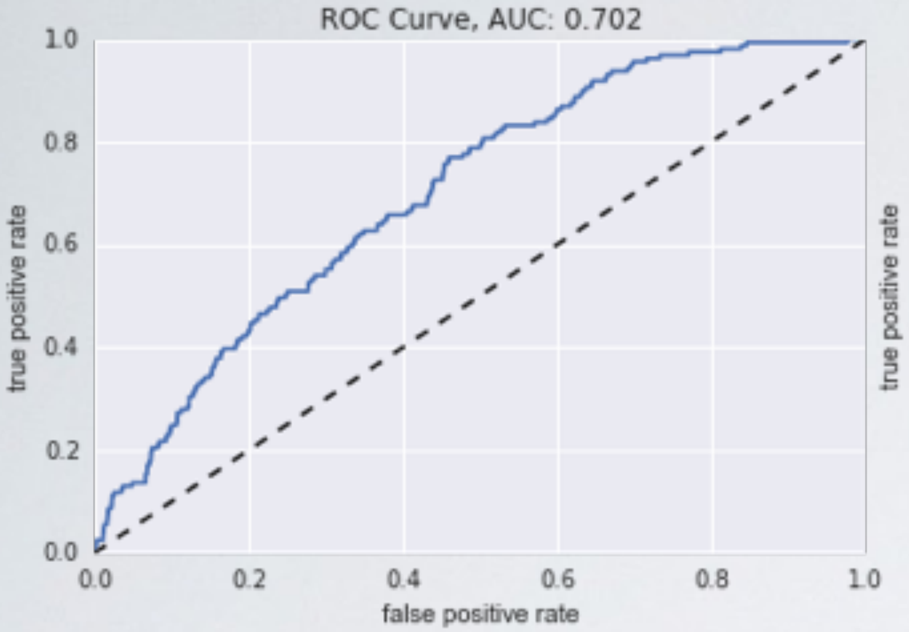


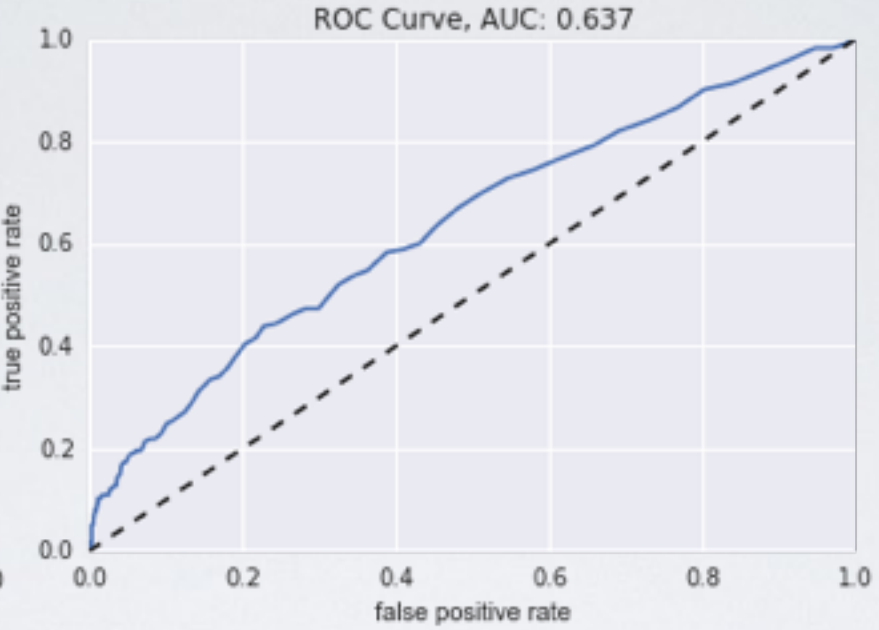
"Live attributions" excludes attributions after signup date

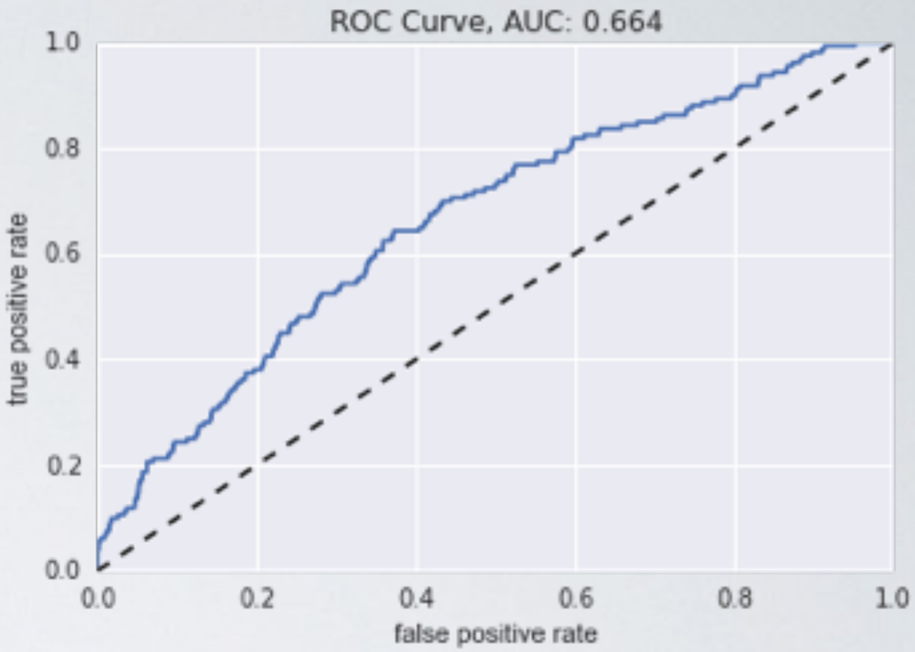
MODEL VARIATIONS

Model	Features	Transformations
Logistic Regression (L1 Regularization)	All phone data	Encode unique daily attribution
Random Forests	Initial phone	Feature importance > 0.001
Gradient Boosted Trees	Prior to initial phone	
	No phone data	
	No attribution	

CROSS VALIDATION AND MODEL CHOICE







Top decile ratio*: 2.167

Gradient Boosted Trees:
n_estimators = 230
learning_rate = 0.07
max_depth = 5
subsample = 1
max_features = None

2.247

Random Forests:
n_estimators = 300

2.181

Logistic Regression:
Penalty = L1

* top decile signup rate / average signup rate

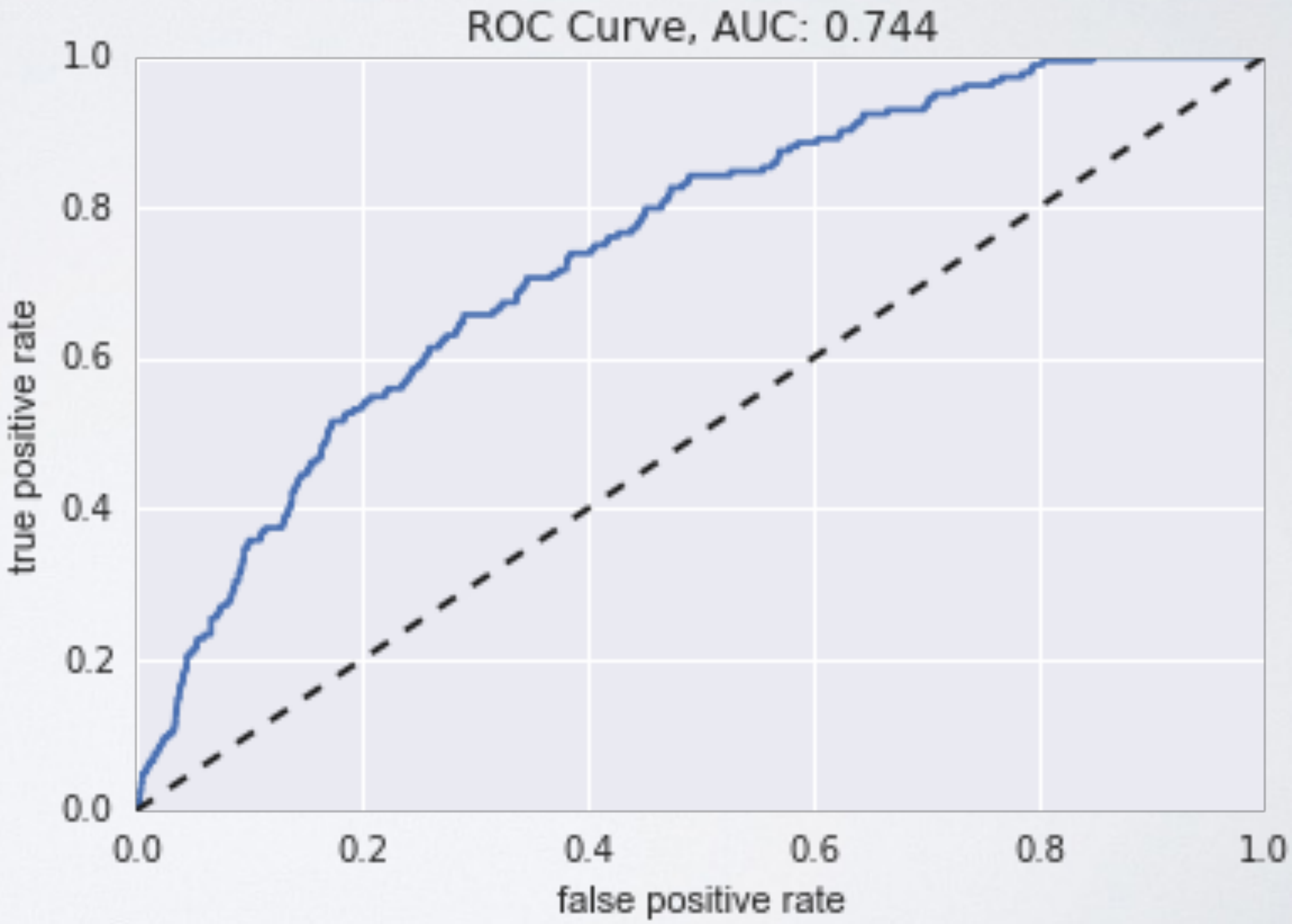
BEST RESULT: GB TREES, ALL DATA

top decile signup rate: 0.2103
average signup rate: 0.0732
top decile ratio: 2.8713

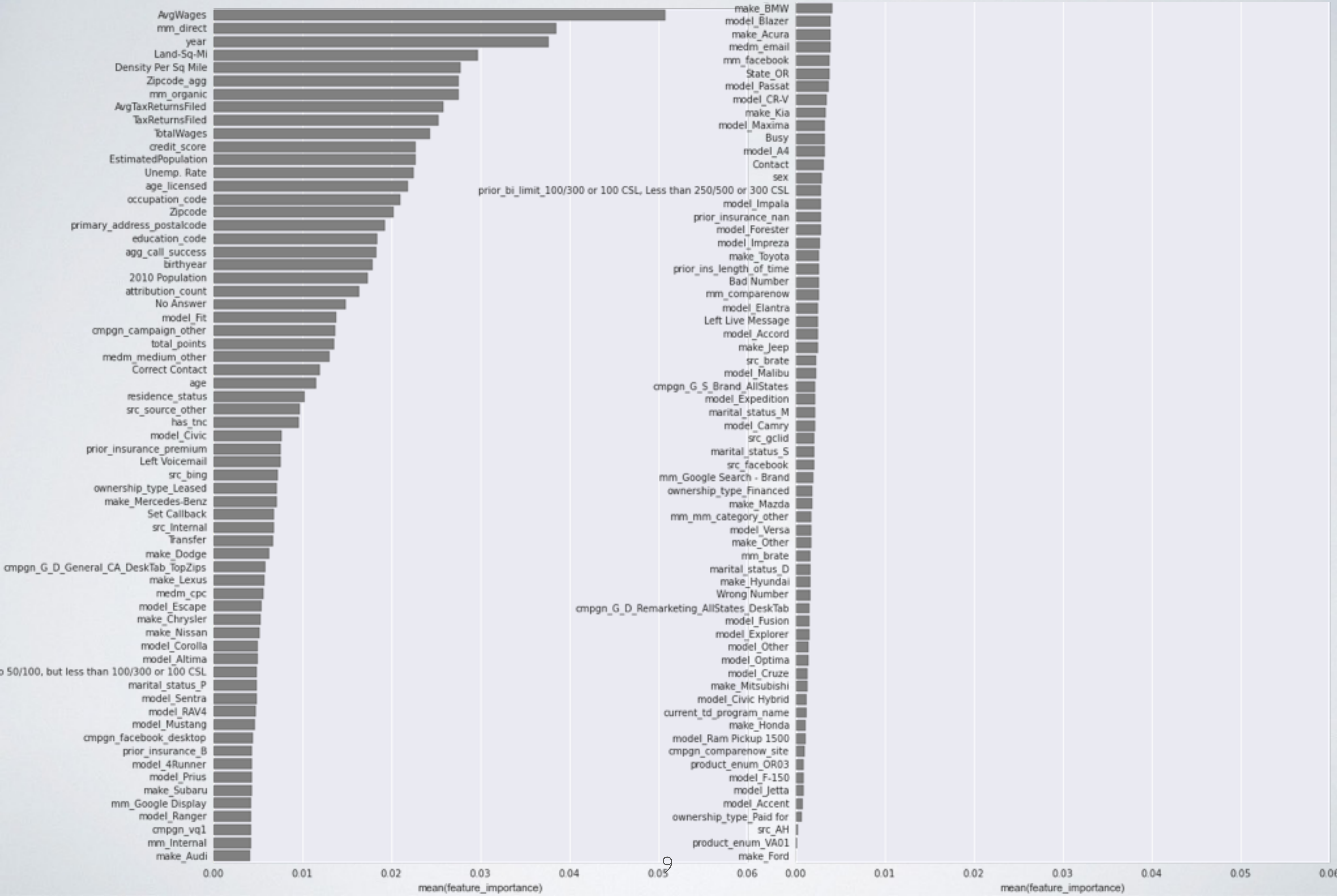
F1 Score: 0.069
accuracy: 0.9248
precision: 0.3684
recall: 0.038

confusion matrix:

```
   Pred F   Pred T
F  [2316    12]
T  [ 177     7]
```



GB TREE FEATURE IMPORTANCE



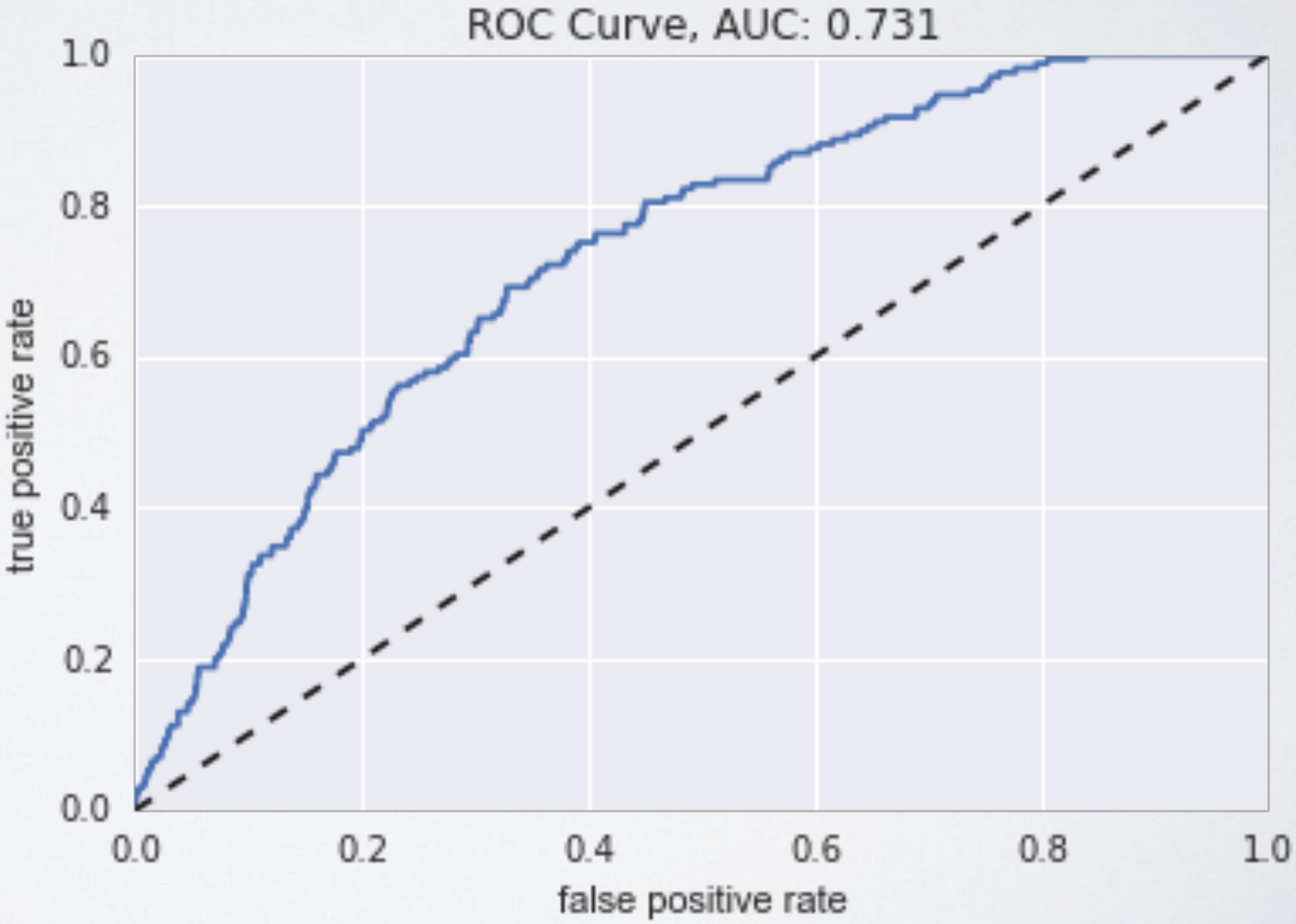
INITIAL PHONE CALL

top decile signup rate: 0.1667
average signup rate: 0.0673
top decile ratio: 2.4773

F1 Score: 0.0529
accuracy: 0.9287
precision: 0.2500
recall: 0.0296

confusion matrix:

```
   Pred F   Pred T
F   [2328    15]
T   [ 164     5]
```



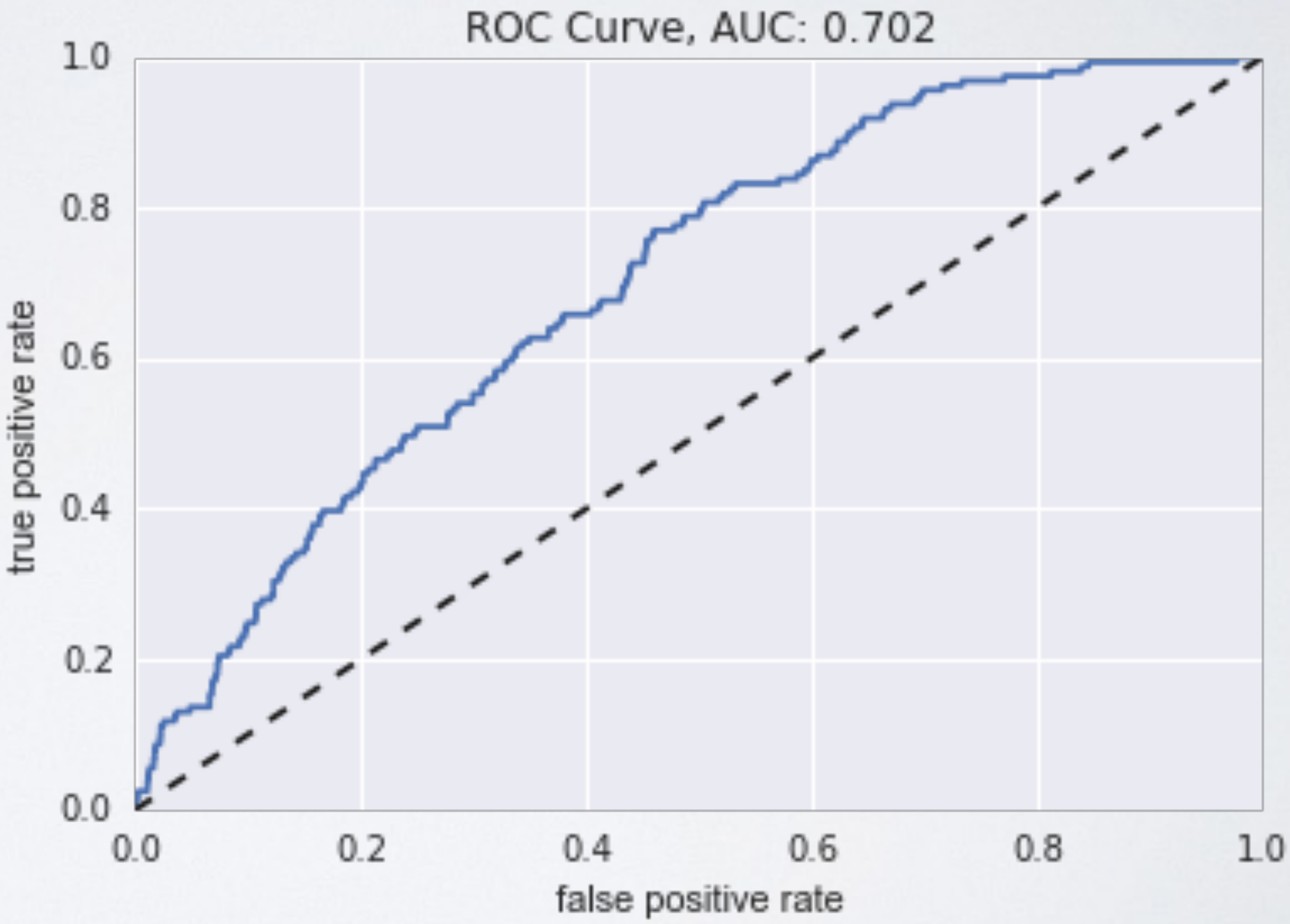
PRIOR TO INITIAL CALL*

top decile signup rate: 0.1389
average signup rate: 0.0641
top decile ratio: 2.167

F1 Score: 0.0455
accuracy: 0.9331
precision: 0.2667
recall: 0.0248

confusion matrix:

	Pred F	Pred T
F	[2340	11]
T	[157	4]



* Excludes call result (e.g. voicemail, no answer, transfer)

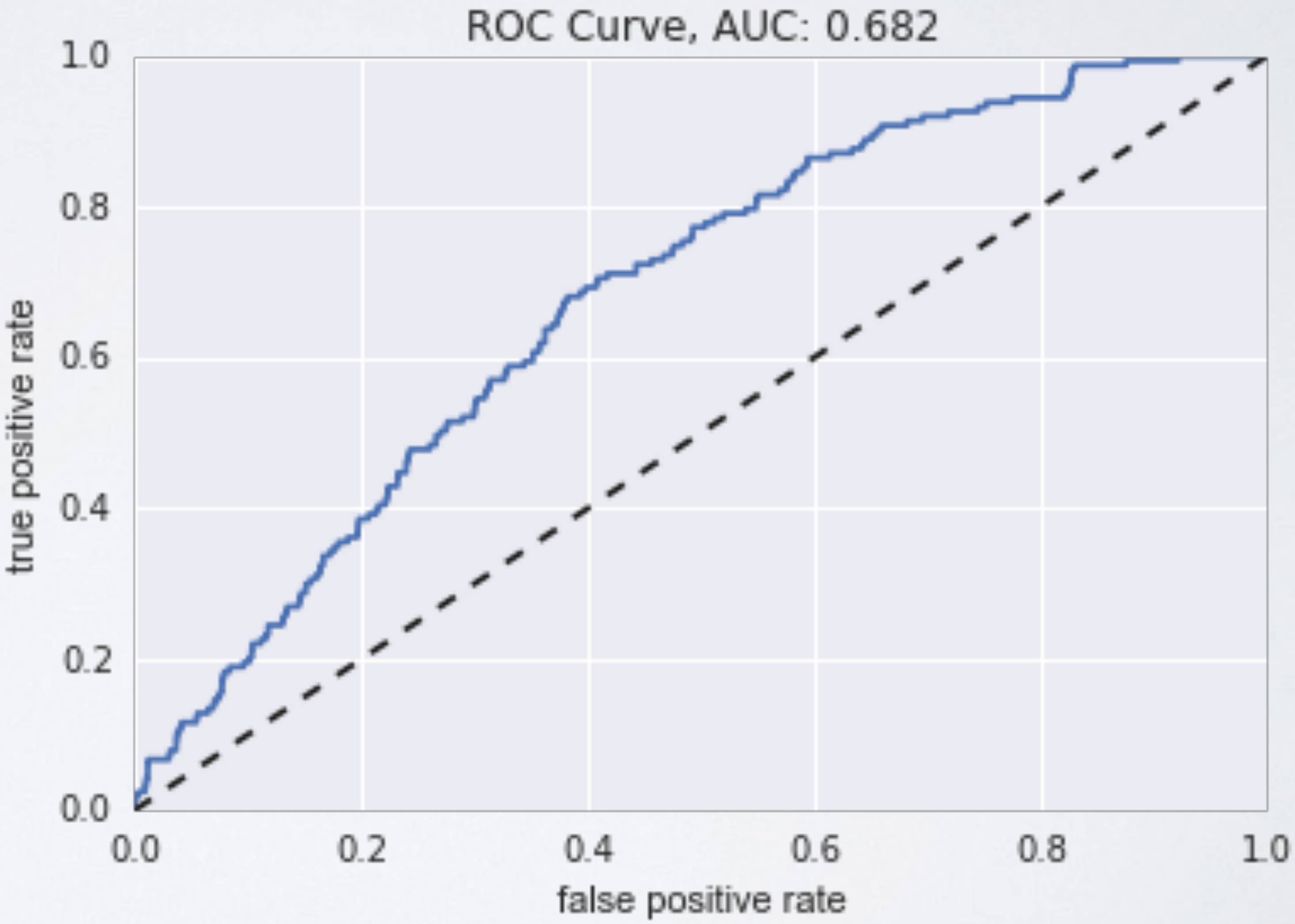
EXCLUDE ALL PHONE DATA

top decile signup rate: 0.1230
average signup rate: 0.0649
top decile ratio: 1.8958

F1 Score: 0.0435
accuracy: 0.9299
precision: 0.1905
recall: 0.0245

confusion matrix:

```
   Pred F  Pred T
F  [2332    17]
T  [ 159     4]
```



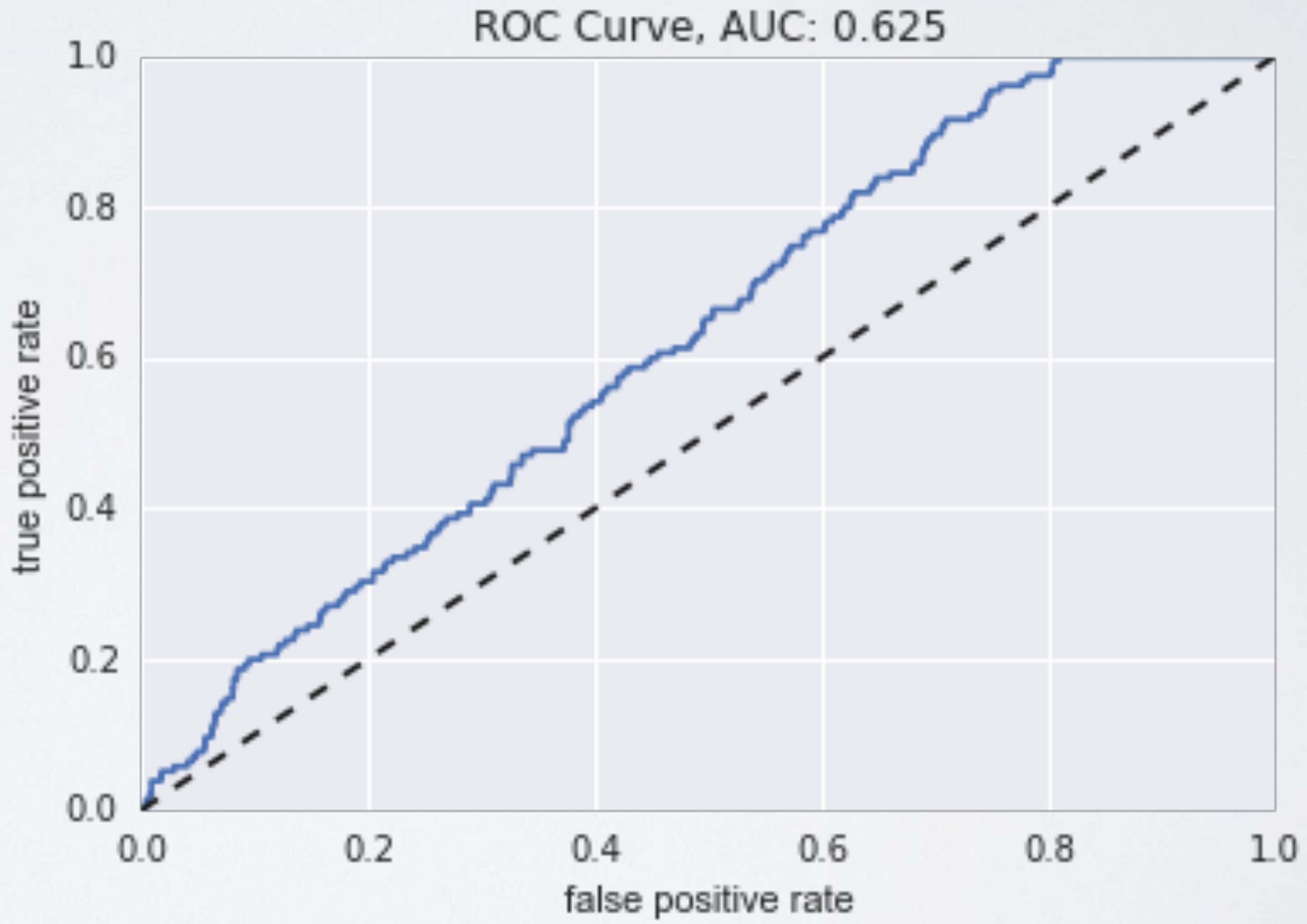
INITIAL PHONE, EXCLUDE ALL ATTRIBUTION DATA

top decile signup rate: 0.0119
average signup rate: 0.0617
top decile ratio: 1.9293

F1 Score: 0.0120
accuracy: 0.9347
precision: 0.0909
recall: 0.0065

confusion matrix:

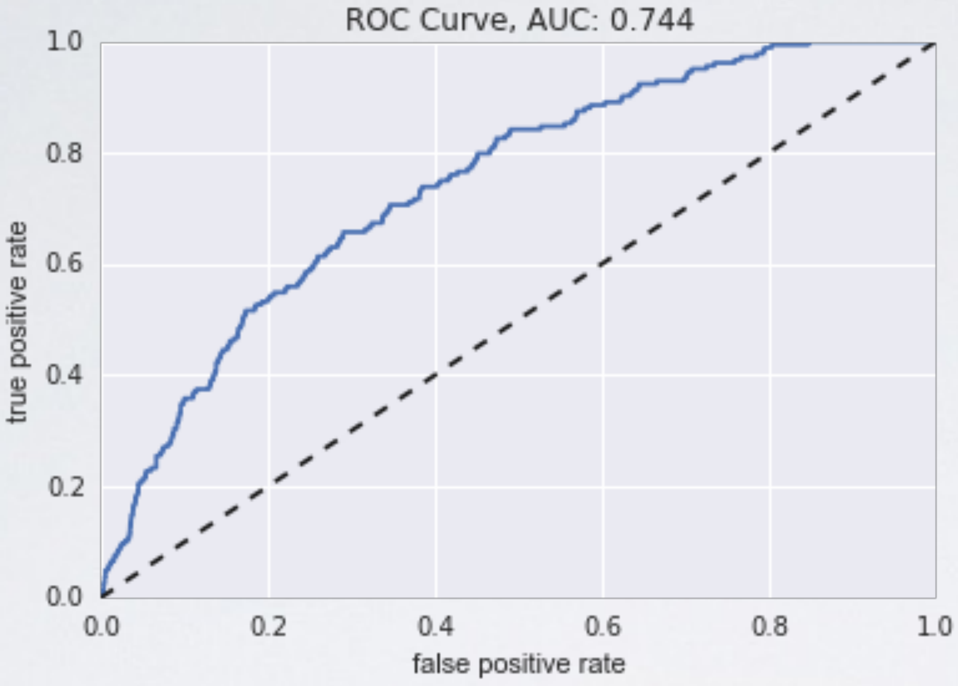
```
   Pred F  Pred T
F  [2342    10]
T  [ 149     1]
```



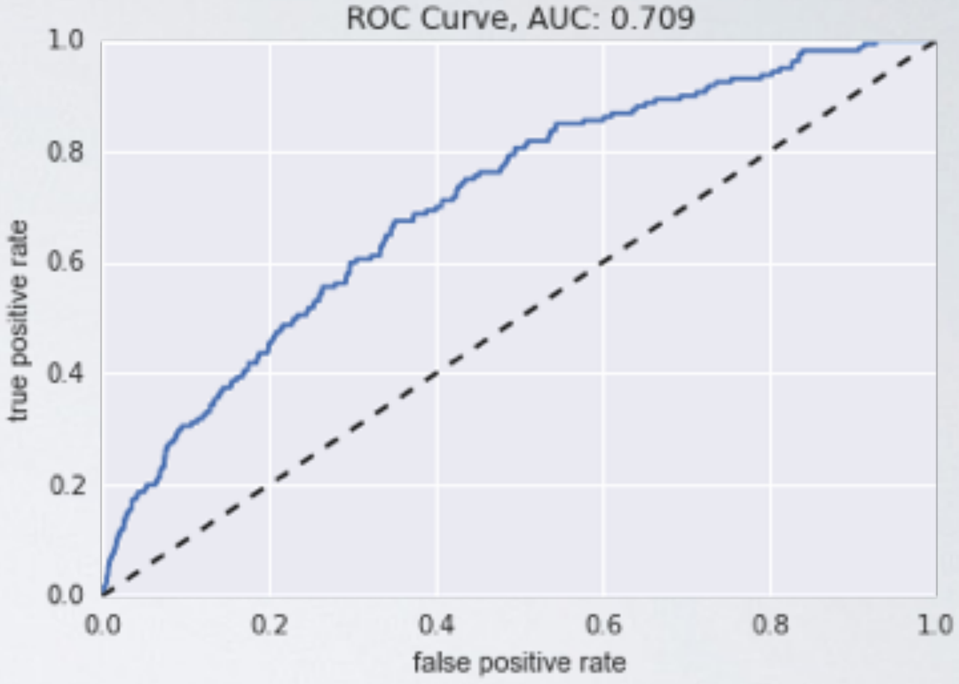
ENCODING ATTRIBUTION: UNIQUE DAILY ACTIVITY

Base Case **Encoded Attribution**

All phone data

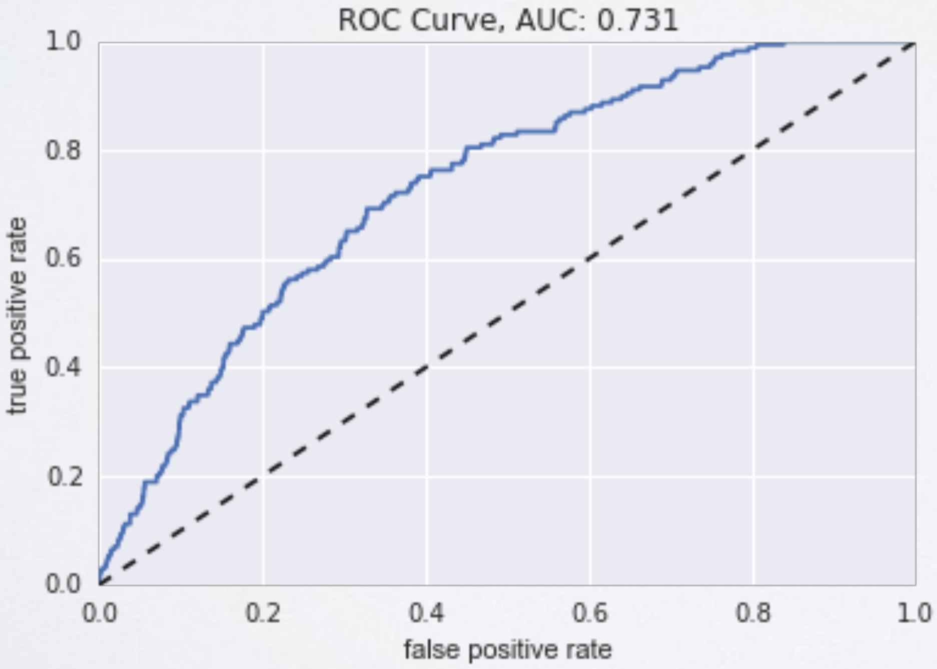


Top decile ratio: 2.8713

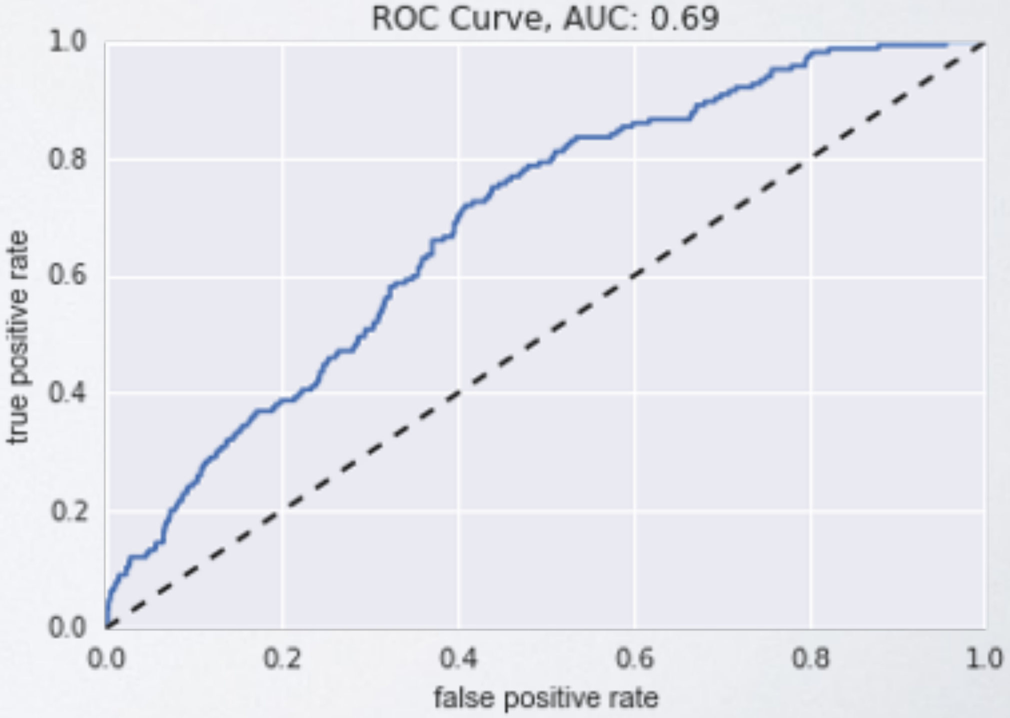


2.8036

Initial phone data

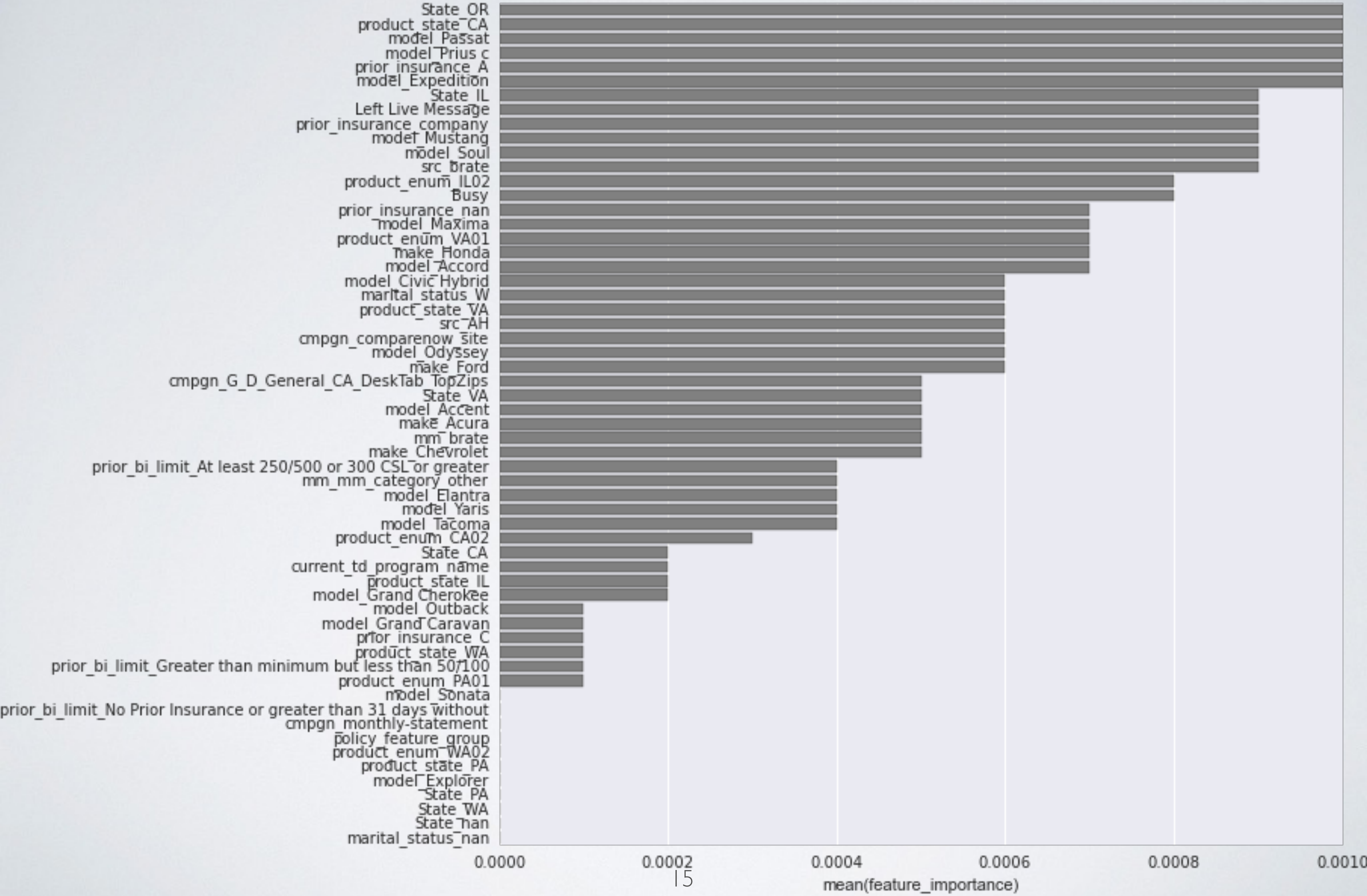


Top decile ratio: 2.4773



2.2957

FEATURE IMPORTANCE <0.001



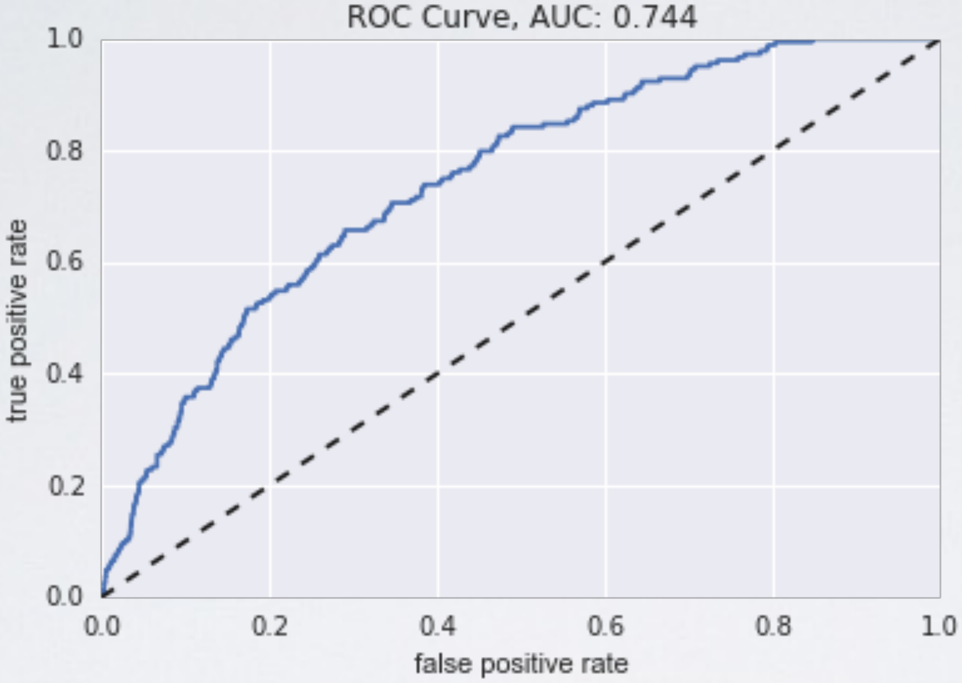
SELECTING FEATURE IMPORTANCE >0.001

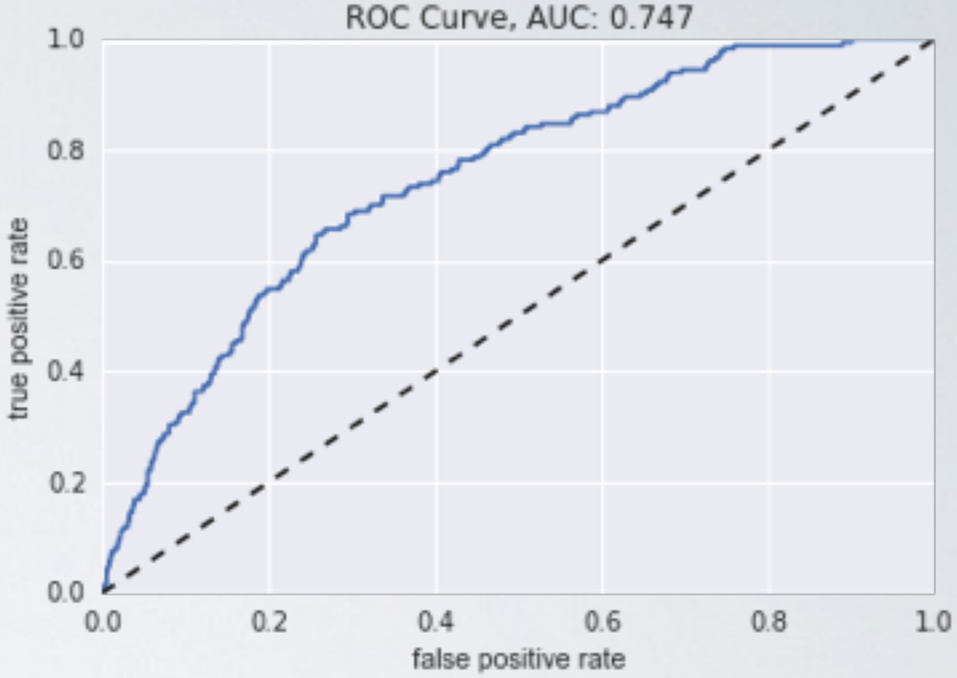
Base Case

Selected Features

All phone data

features 184 -> 134



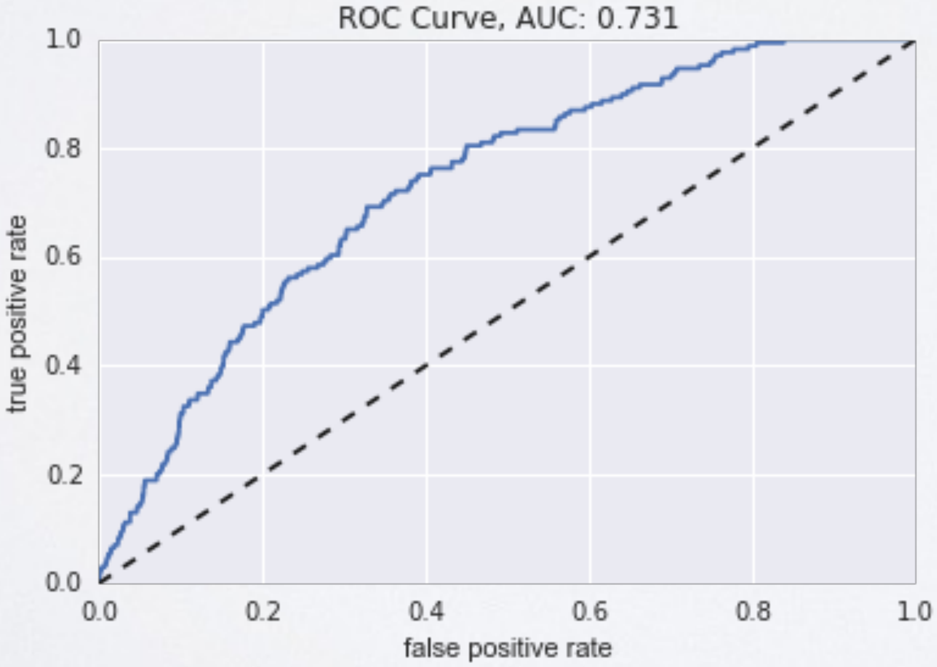


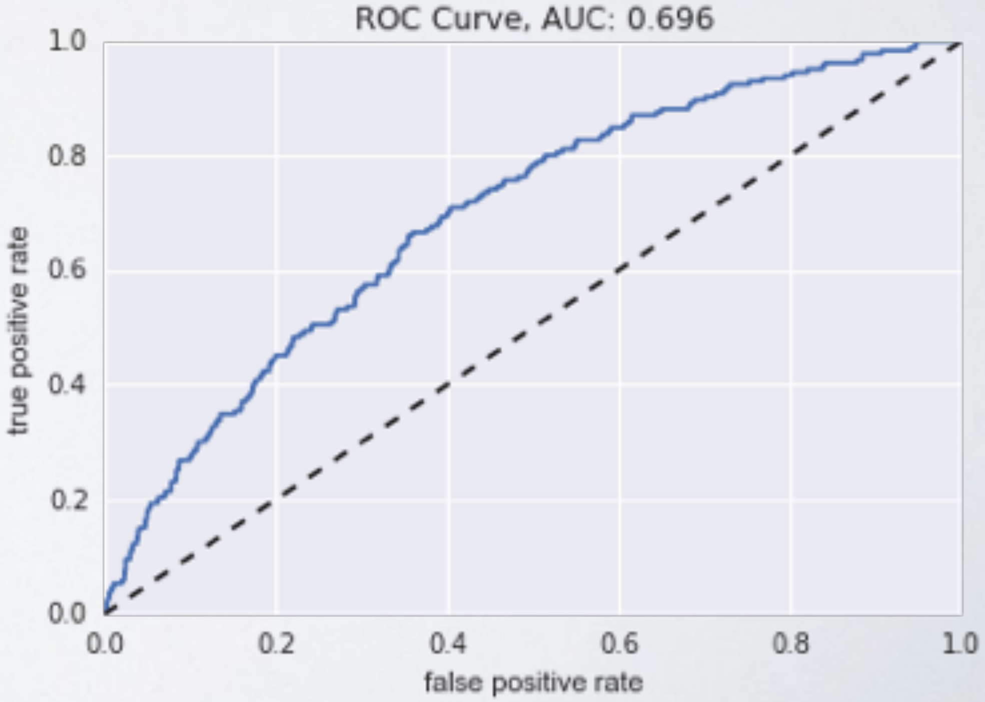
Top decile ratio: 2.8713

3.0338

Initial phone data

features 190 -> 133





Top decile ratio: 2.4773

2.8902

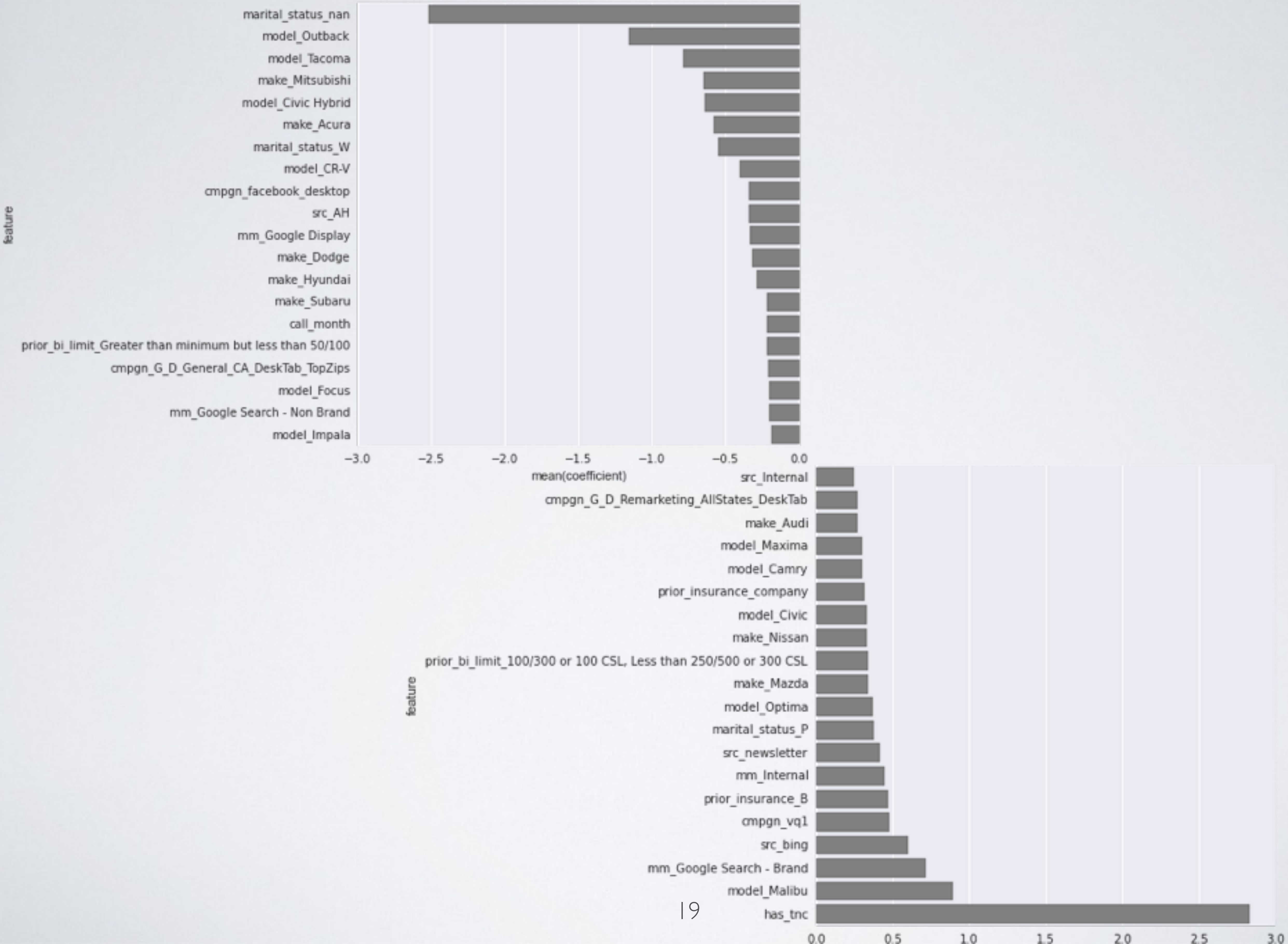
FINDINGS SUMMARY

- Model prediction of top decile is 2.5x of overall signup rate
- GB Trees, Random Forests, and Logistic models are predictive
- Online attribution data improves prediction
- Phone data also improves prediction. Aggregate call success and call results (voicemail, no answer, etc.)
- Filtering feature importance >0.001 reduces number of features by 1/3rd without impacting prediction.

Recommendation: Add online attribution and zip code data to the phone team's lead scoring model

APPENDIX

TOP AND BOTTOM 20 LOGISTIC COEFFICIENTS



CONTINUING WORK

- Average top decile numbers to get more stable assessments.
- Incorporate secondary driver information

OTHER THINGS ANDREW TRIED

- SVM. didn’t produce better results than GBTrees. Given time, didn’t invest much time tuning hyper parameters

- XGBoost. Relatively small data set, so didn’t need the speed. More familiar with sclera

53 FEATURES EXCLUDED, FEATURE IMPORTANCE < 0.001

```
'marital_status_W',
 'marital_status_nan',
 'State_CA',
 'State_IL',
 'State_PA',
 'State_VA',
 'State_WA',
 'State_nan',
 'policy_feature_group',
 'prior_insurance_company',
 'prior_liability_c',
 'product_state_CA',
 'product_state_IL',
 'product_state_OR',
 'product_state_PA',
 'product_state_VA',
 'product_state_WA',
 'product_enum_CA02',
 'product_enum_IL02',
 'product_enum_PA01',
 'product_enum_WA02',
 'prior_insurance_A',
 'prior_insurance_C',
 'prior_bi_limit_At least 250/500 or 300 CSL or greater',
 'prior_bi_limit_Greater than minimum but less than 50/100',
 'prior_bi_limit_No Prior Insurance or greater than 31 days without',
'make_Chevrolet',
'make_Volkswagen',
'model_328',
'model_Focus',
'model_Grand Caravan',
'model_Grand Cherokee',
'model_MAZDA3',
'model_Odyssey',
'model_Outback',
'model_Prius c',
'model_Sienna',
'model_Silverado 1500',
'model_Sonata',
'model_Soul',
'model_Tacoma',
'model_Taurus',
'model_Yaris',
'mm_Google Search - Non Brand',
'mm_QL',
'src_QL',
'src_UE',
'src_comparenow',
'src_newsletter',
'cmpgn_QL',
'cmpgn_brate',
'cmpgn_monthly-statement',
'cmpgn_nanigans_mobile'
```

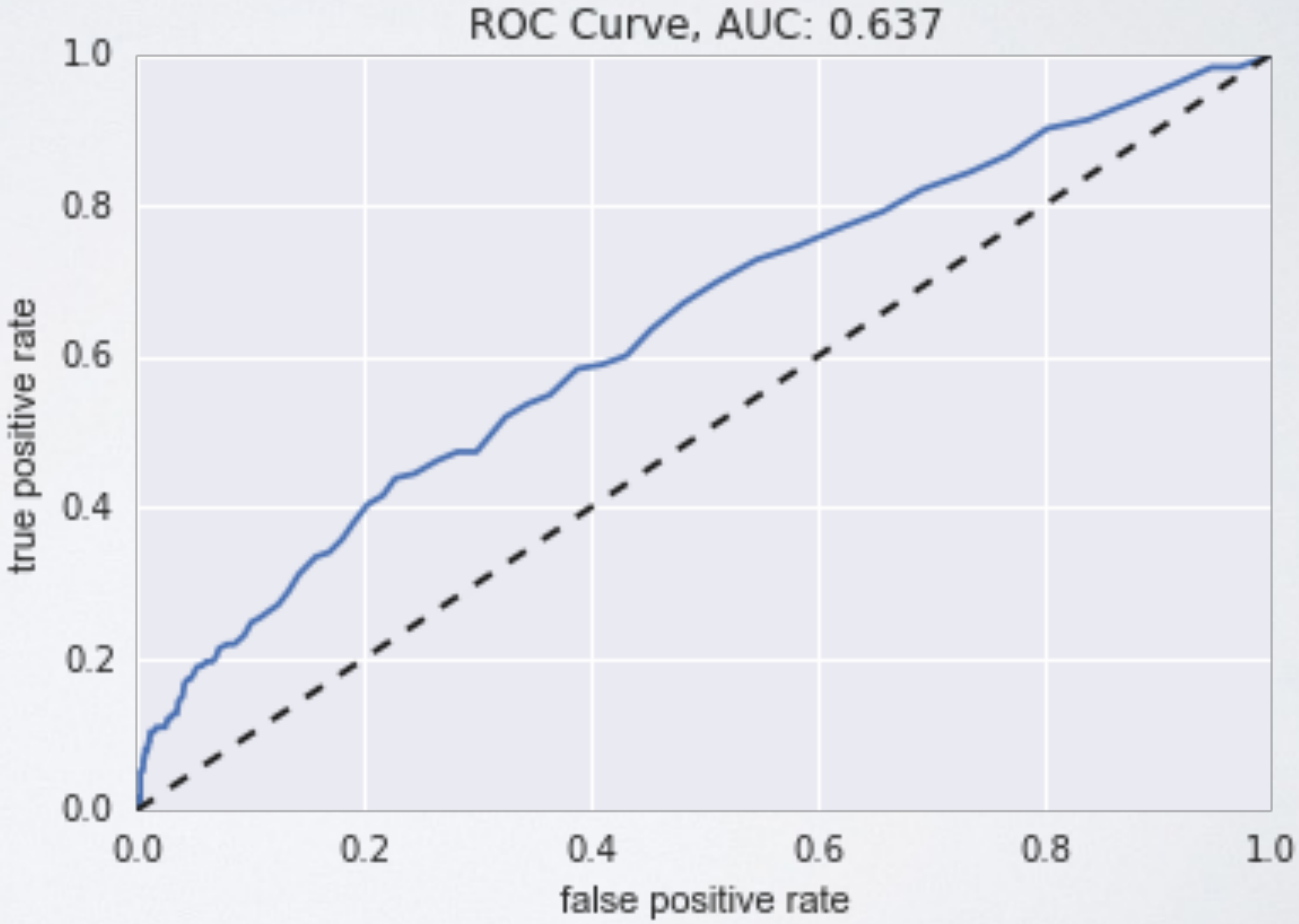
PRIOR TO INITIAL CALL, RANDOM FORESTS*

top decile esign rate: 0.1548
average esign rate: 0.0689
top decile esign / avg. esign: 2.2472

F1 Score: 0.0000
accuracy: 0.9311
precision: 0.0000
recall: 0.0000

confusion matrix:

```
   Pred F  Pred T
F  [2339     0]
T  [ 173     0]
```



*300 estimators

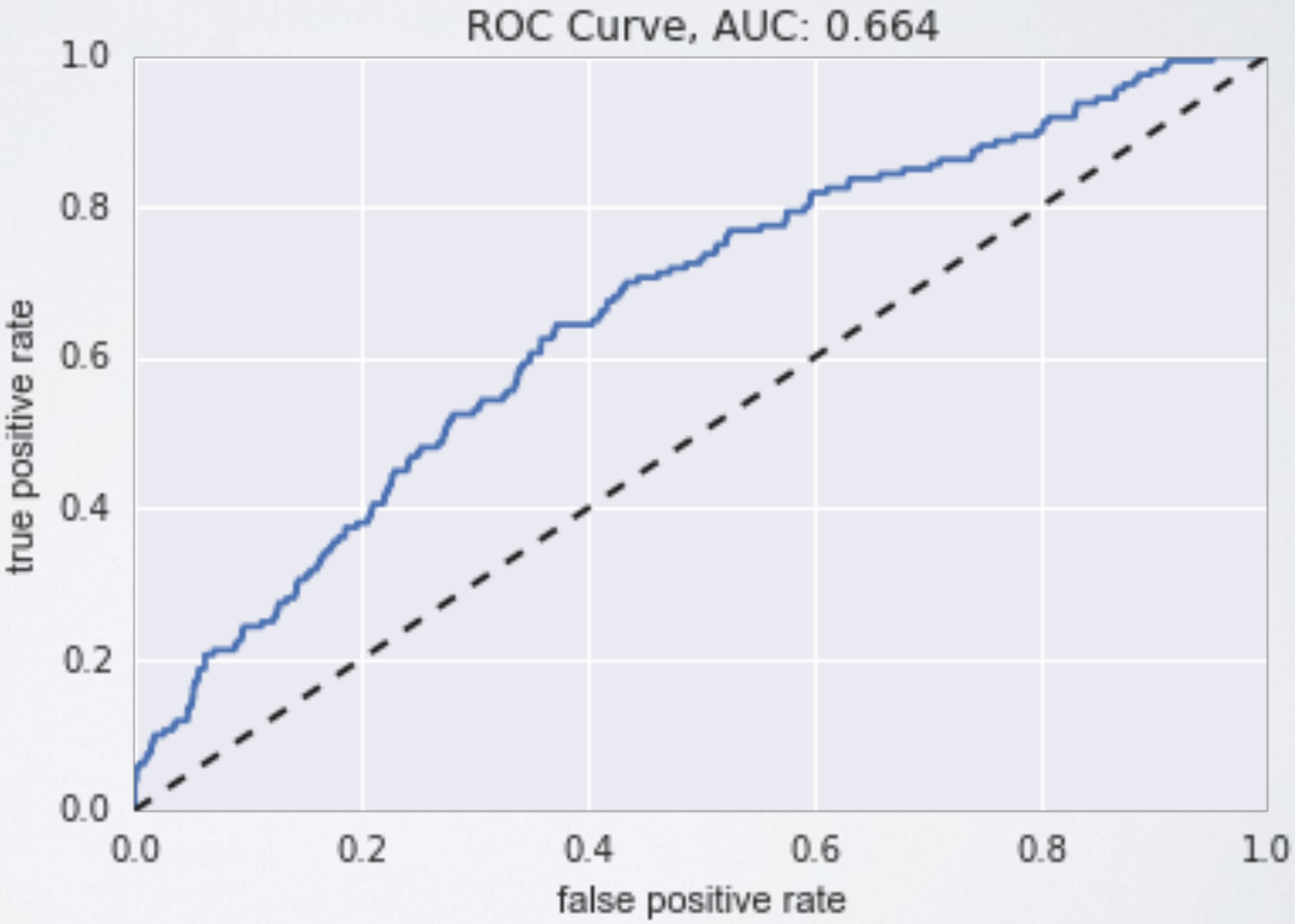
PRIOR TO INITIAL CALL, L1 LOGISTIC REGRESSION

top decile esign rate: 0.1389
average esign rate: 0.0637
top decile esign / avg. esign: 2.1806

F1 Score: 0.0819
accuracy: 0.9375
precision: 0.6364
recall: 0.0437

confusion matrix:

	Pred F	Pred T
F	[2348	4]
T	[153	7]



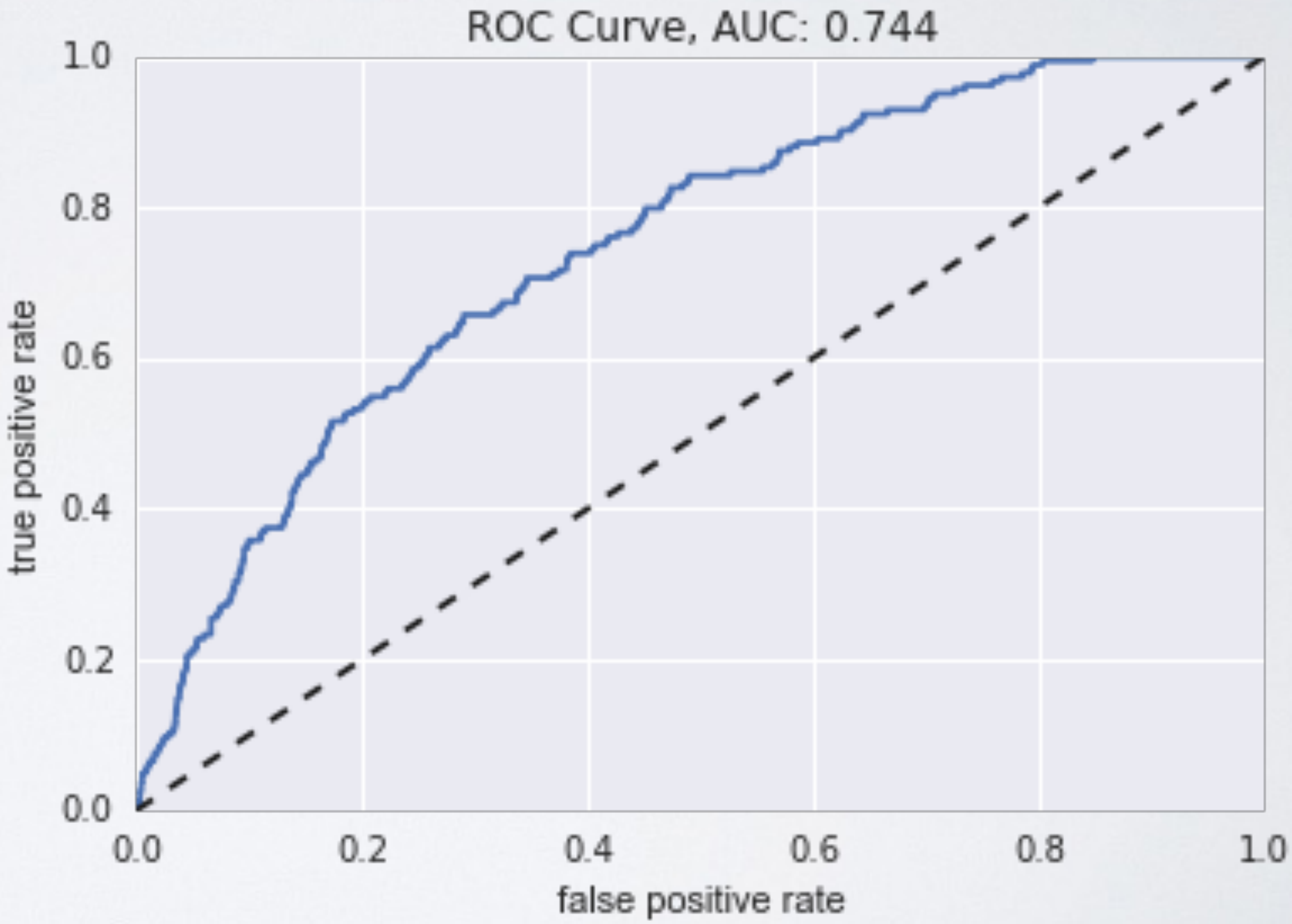
BEST PREDICTIVE MODEL: GB TREES, ALL DATA

top decile esign rate: 0.2103
average esign rate: 0.0732
top decile esign / avg. esign: 2.8713

F1 Score: 0.069
accuracy: 0.9248
precision: 0.3684
recall: 0.038

confusion matrix:

```
    Pred F   Pred T
F   [2316      12]
T   [ 177       7]
```



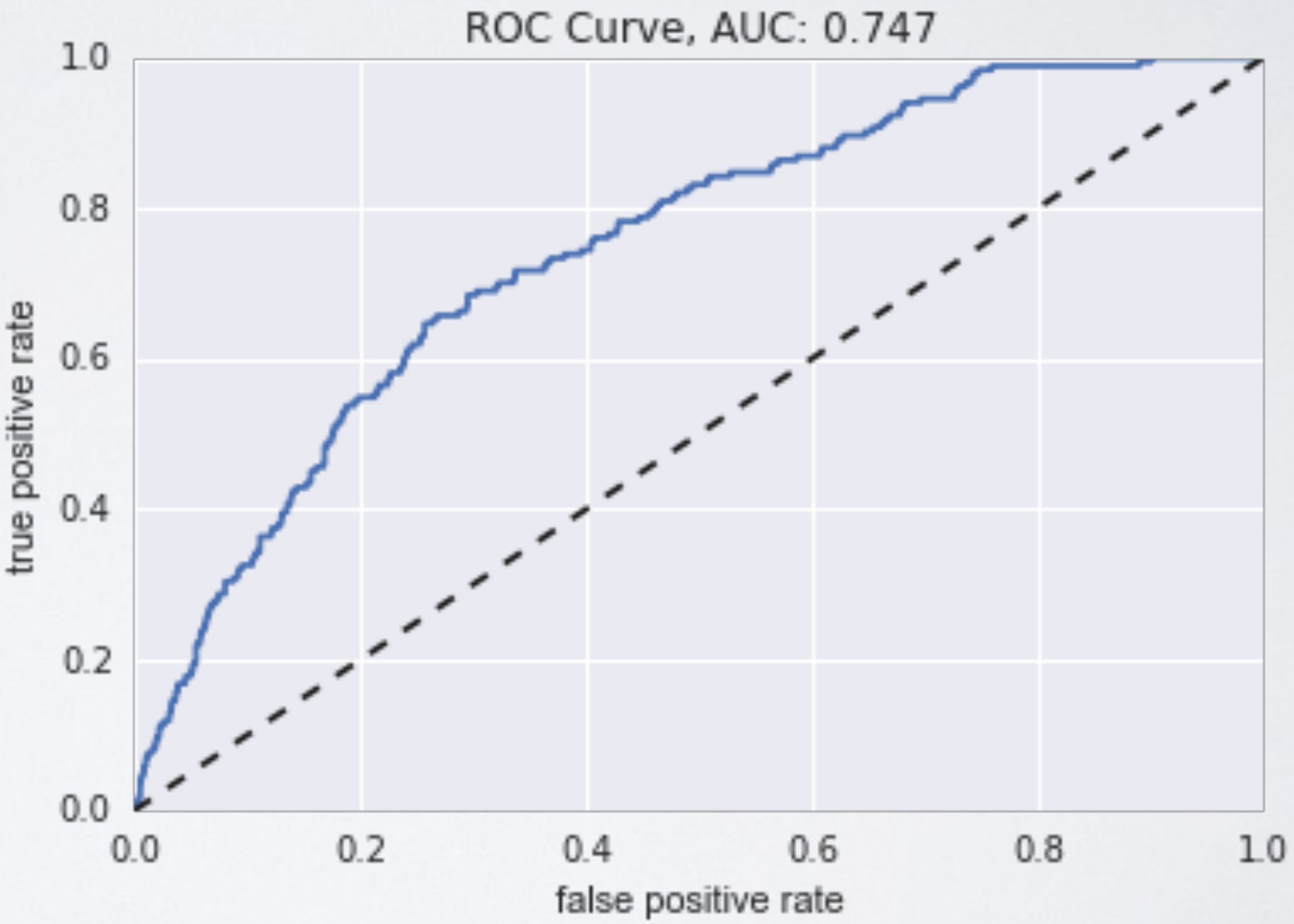
EXCLUDE FEATURE IMPORTANCE < 0.001*

top decile esign rate: 0.2222
average esign rate: 0.0732
top decile esign / avg. esign: 3.0338

F1 Score: 0.0773
accuracy: 0.9240
precision: 0.3478
recall: 0.0435

confusion matrix:

```
   Pred F  Pred T
F  [2313    15]
T  [ 176     8]
```



* Mostly make, model, policy state (full list in appendix)

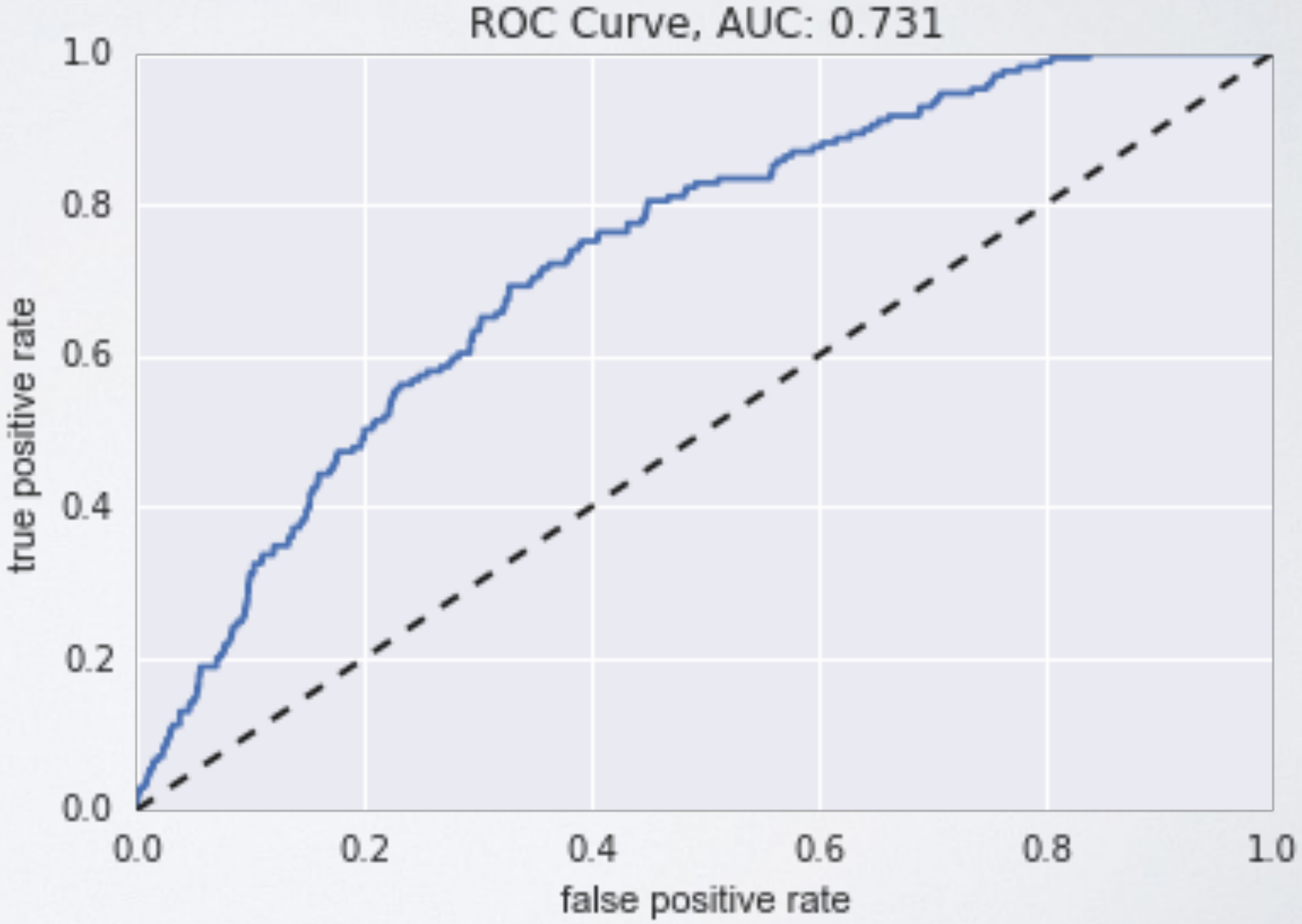
INITIAL PHONE CALL DATA

top decile esign rate: 0.1667
average esign rate: 0.0673
top decile esign / avg. esign: 2.4773

F1 Score: 0.0529
accuracy: 0.9287
precision: 0.2500
recall: 0.0296

confusion matrix:

```
   Pred F  Pred T
F  [2328   15]
T  [ 164    5]
```



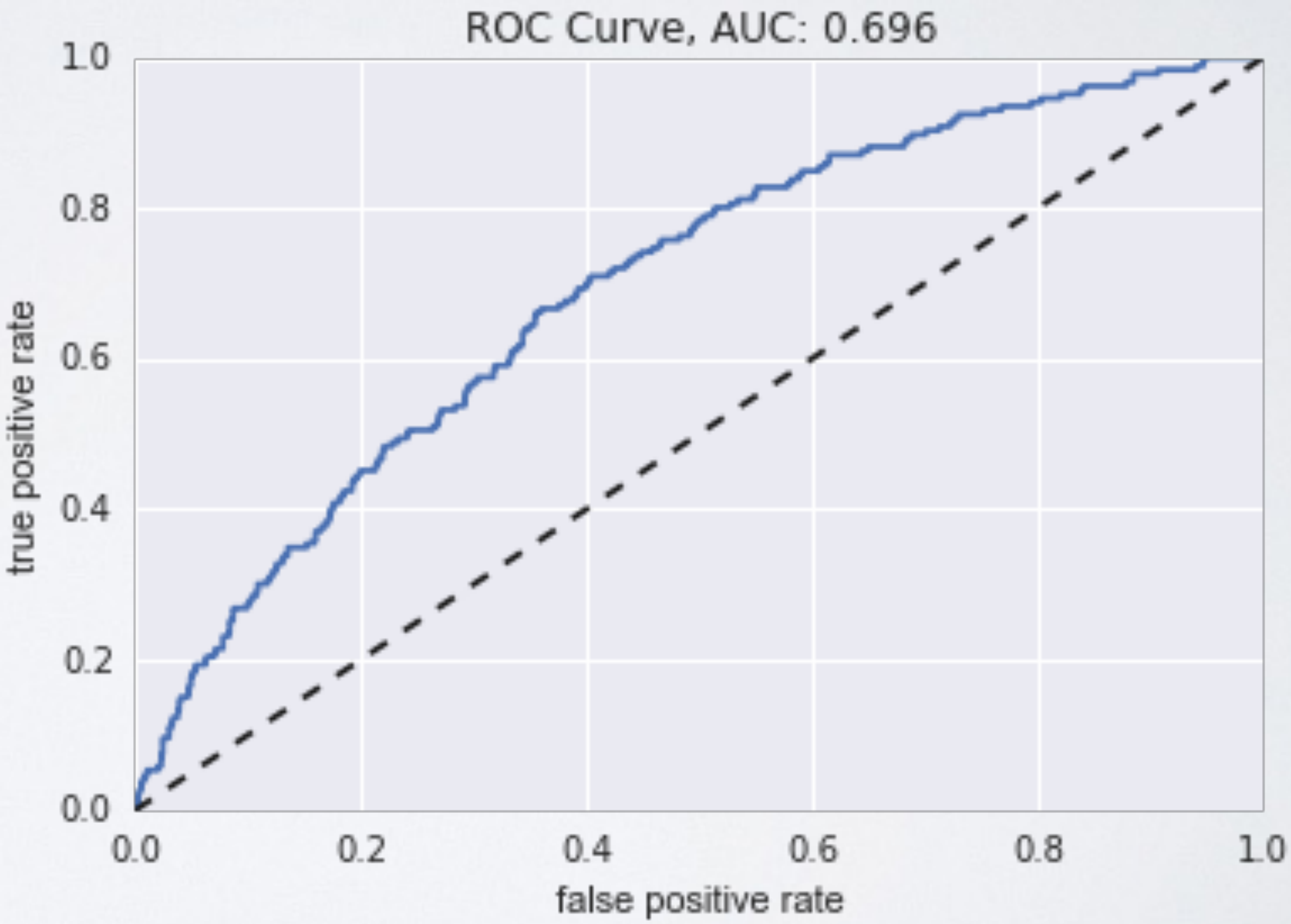
INITIAL PHONE CALL, FEATURE IMPORTANCE > 0.001*

top decile esign rate: 0.1944
average esign rate: 0.0673
top decile esign / avg. esign: 2.8902

F1 Score: 0.0521
accuracy: 0.9275
precision: 0.2174
recall: 0.0296

confusion matrix:

	Pred F	Pred T
F	[2325	18]
T	[164	5]



* 190 features reduced to 133

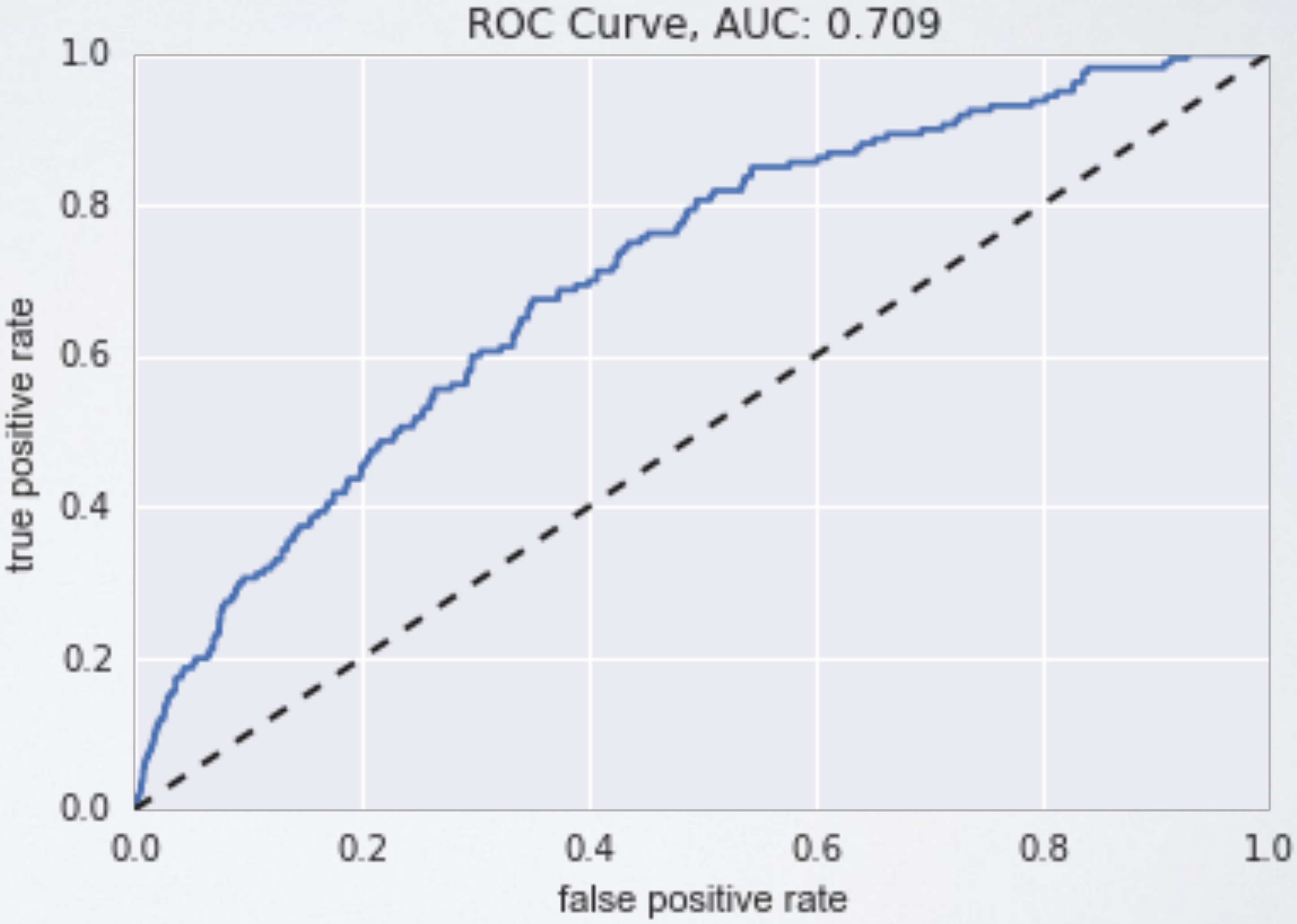
TRANSFORMED ATTRIBUTION TO UNIQUE DAILY VISITS*

top decile esign rate: 0.1796
average esign rate: 0.0637
top decile esign / avg. esign: 2.8036

F1 Score: 0.0757
accuracy: 0.9319
precision: 0.2800
recall: 0.0437

confusion matrix:

```
   Pred F   Pred T
F   [2334    18]
T   [ 153     7]
```



* Includes both original and transformed attribution features

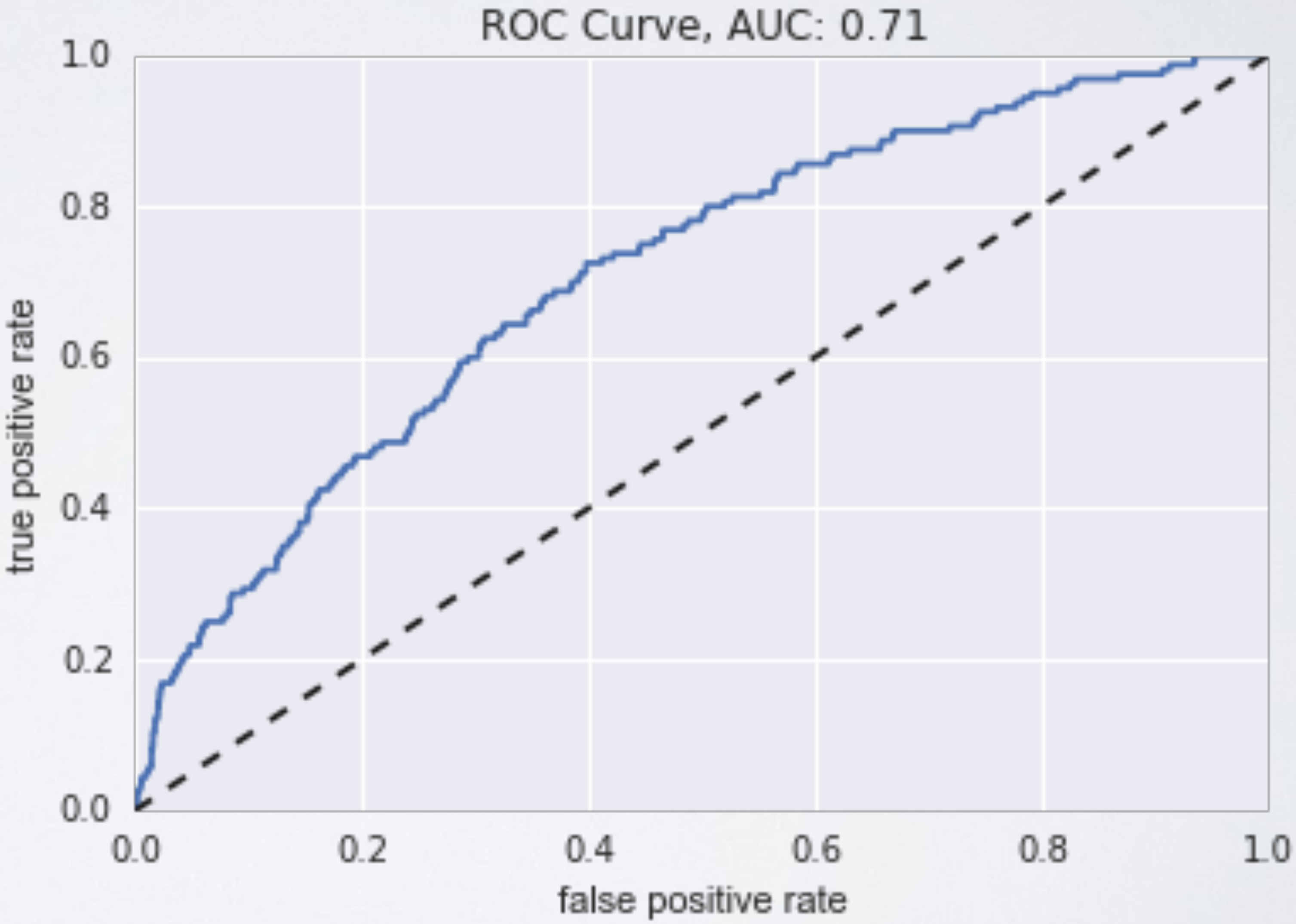
TRANSFORMED ATTRIBUTION, EXCLUDED FEATURE IMPORTANCE < 0.001*

top decile esign rate: 0.1825
average esign rate: 0.0637
top decile esign / avg. esign: 2.8659

F1 Score: 0.0769
accuracy: 0.9331
precision: 0.3182
recall: 0.0437

confusion matrix:

	Pred F	Pred T
F	[2337	15]
T	[153	7]



* 221 to 139

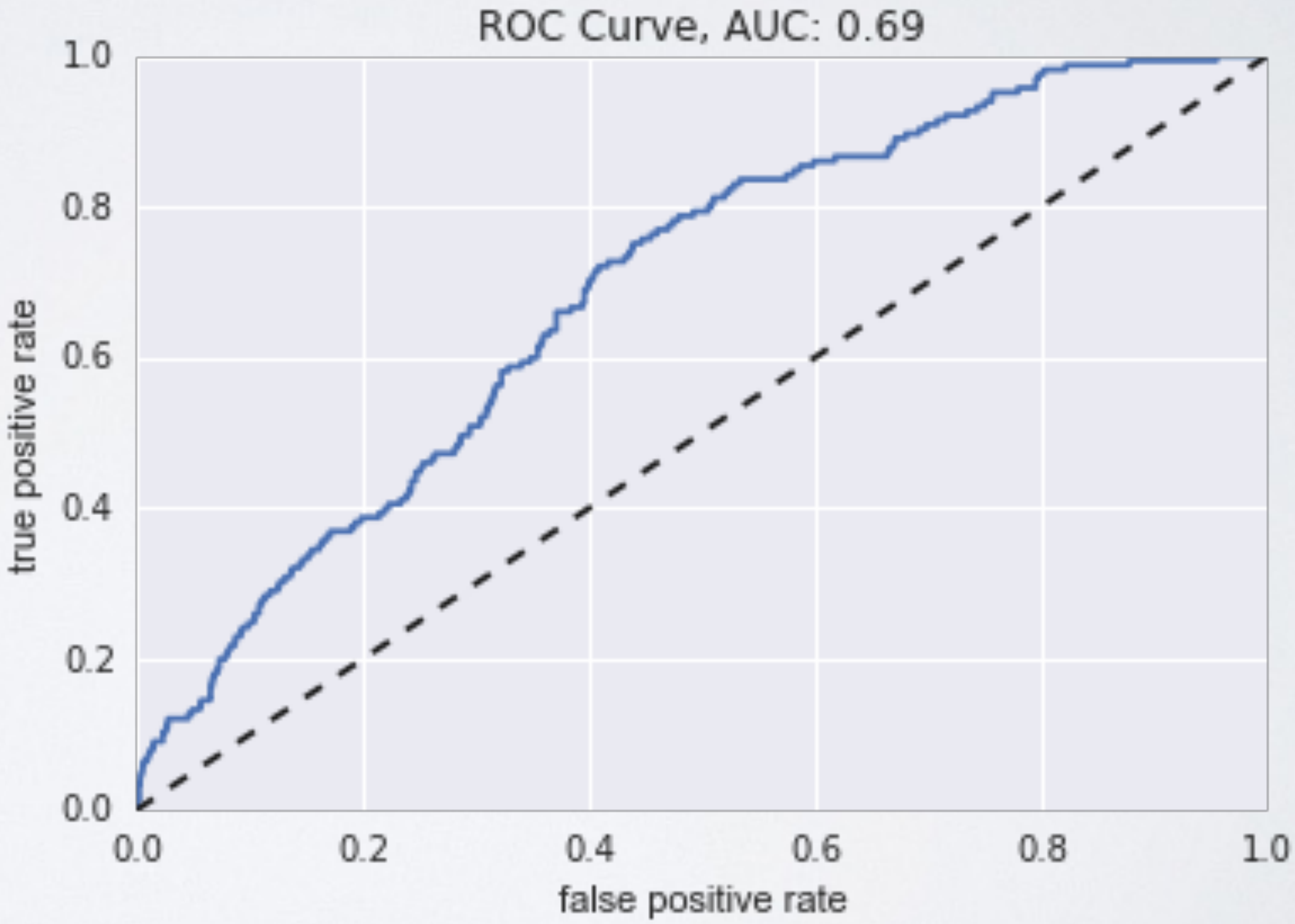
INITIAL PHONE CALL DATA, TRANSFORMED ATTRIBUTION

top decile esign rate: 0.1508
average esign rate: 0.0657
top decile esign / avg. esign: 2.2957

F1 Score: 0.1053
accuracy: 0.9323
precision: 0.4000
recall: 0.0696

confusion matrix:

```
   Pred F  Pred T
F  [2332    15]
T  [ 155    10]
```



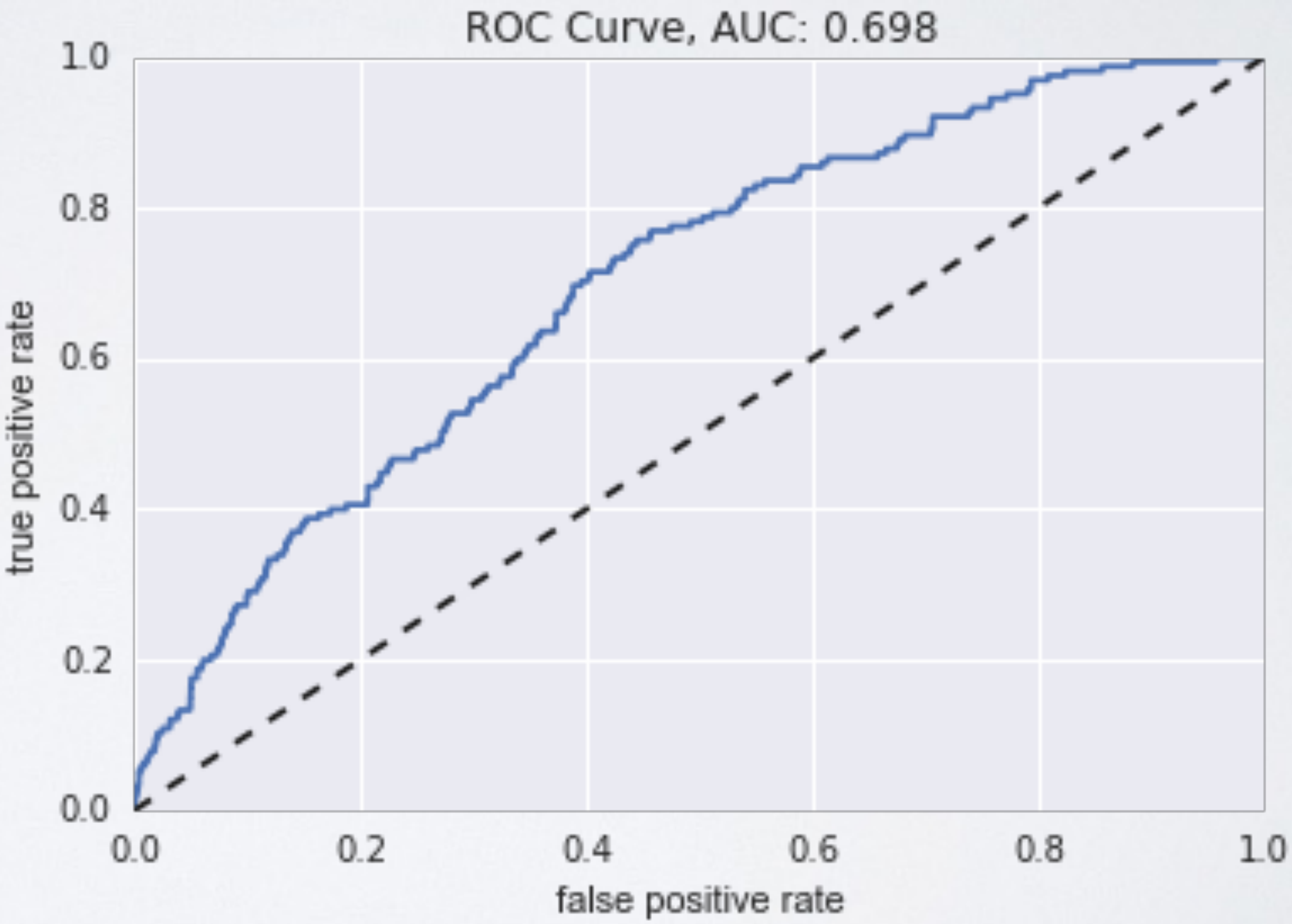
INITIAL PHONE CALL, TRANSFORMED ATTRIBUTION, FEATURE IMPORTANCE > 0.001*

top decile esign rate: 0.1746
average esign rate: 0.0657
top decile esign / avg. esign: 2.6582

F1 Score: 0.0963
accuracy: 0.9327
precision: 0.4091
recall: 0.0545

confusion matrix:

```
   Pred F  Pred T
F  [2334    13]
T  [ 156     9]
```



* 227 features reduced to 138

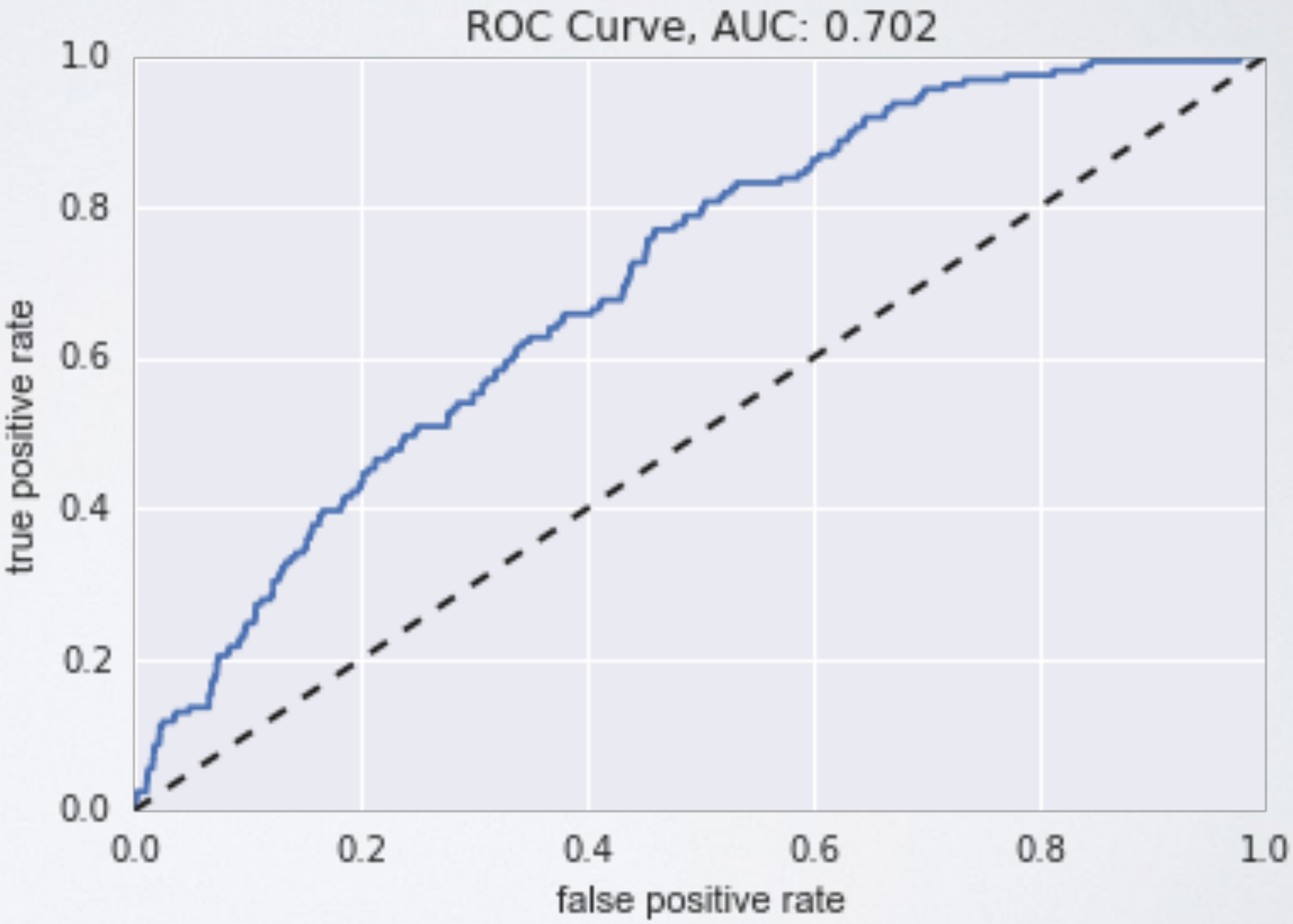
PRIOR TO INITIAL CALL*

top decile esign rate: 0.1389
average esign rate: 0.0641
top decile esign / avg. esign: 2.167

F1 Score: 0.0455
accuracy: 0.9331
precision: 0.2667
recall: 0.0248

confusion matrix:

```
   Pred F  Pred T
F  [2340    11]
T  [ 157     4]
```


ROC Curve, AUC: 0.702

* Excludes call result (e.g. voicemail, no answer, transfer)

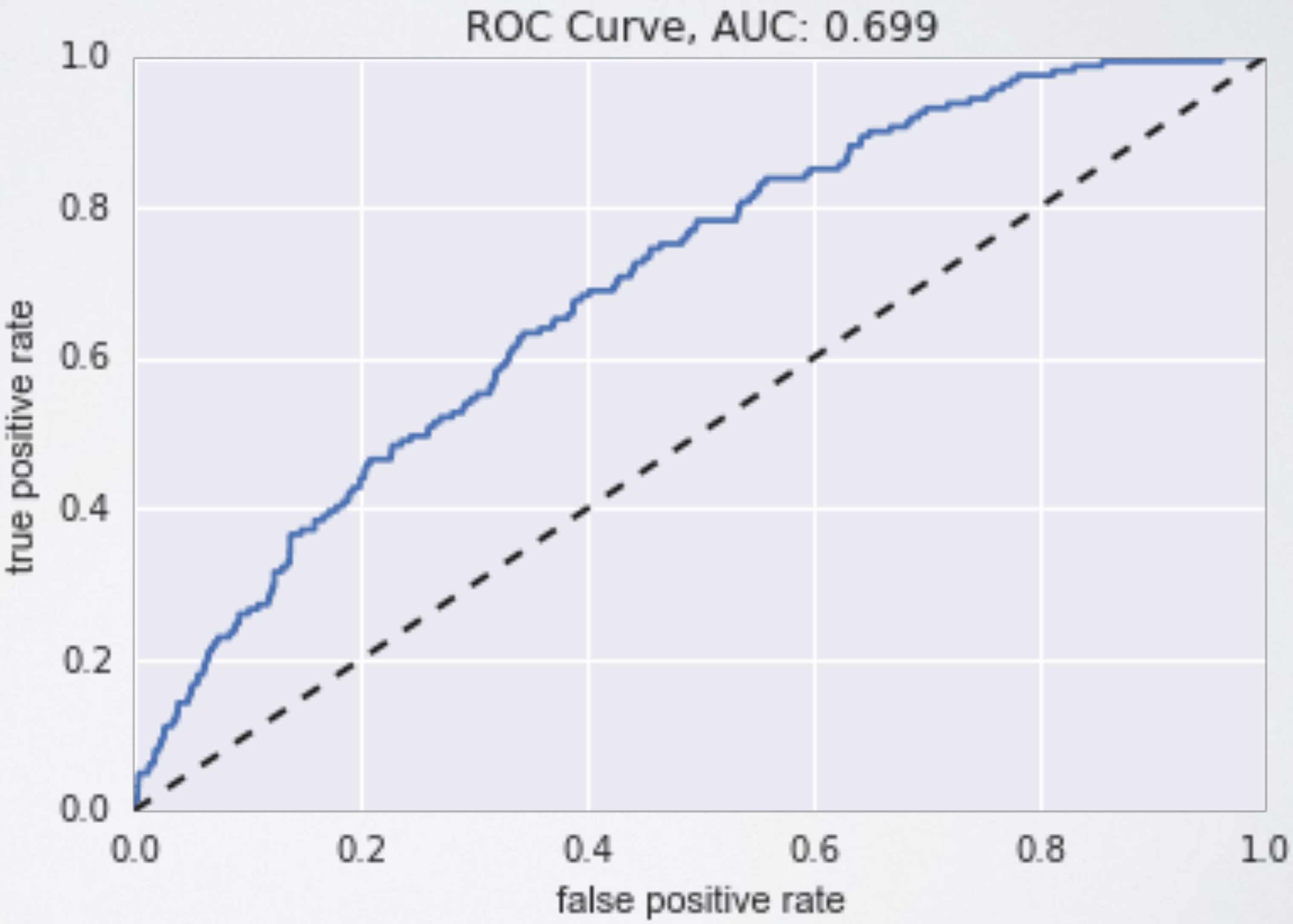
PRIOR TO INITIAL CALL, FEATURE IMPORTANCE > 0.001*

top decile esign rate: 0.1587
average esign rate: 0.0641
top decile esign / avg. esign: 2.4766

F1 Score: 0.0682
accuracy: 0.9347
precision: 0.4000
recall: 0.0373

confusion matrix:

```
   Pred F  Pred T
F  [2342    9]
T  [ 155    6]
```



*178 to 133 features

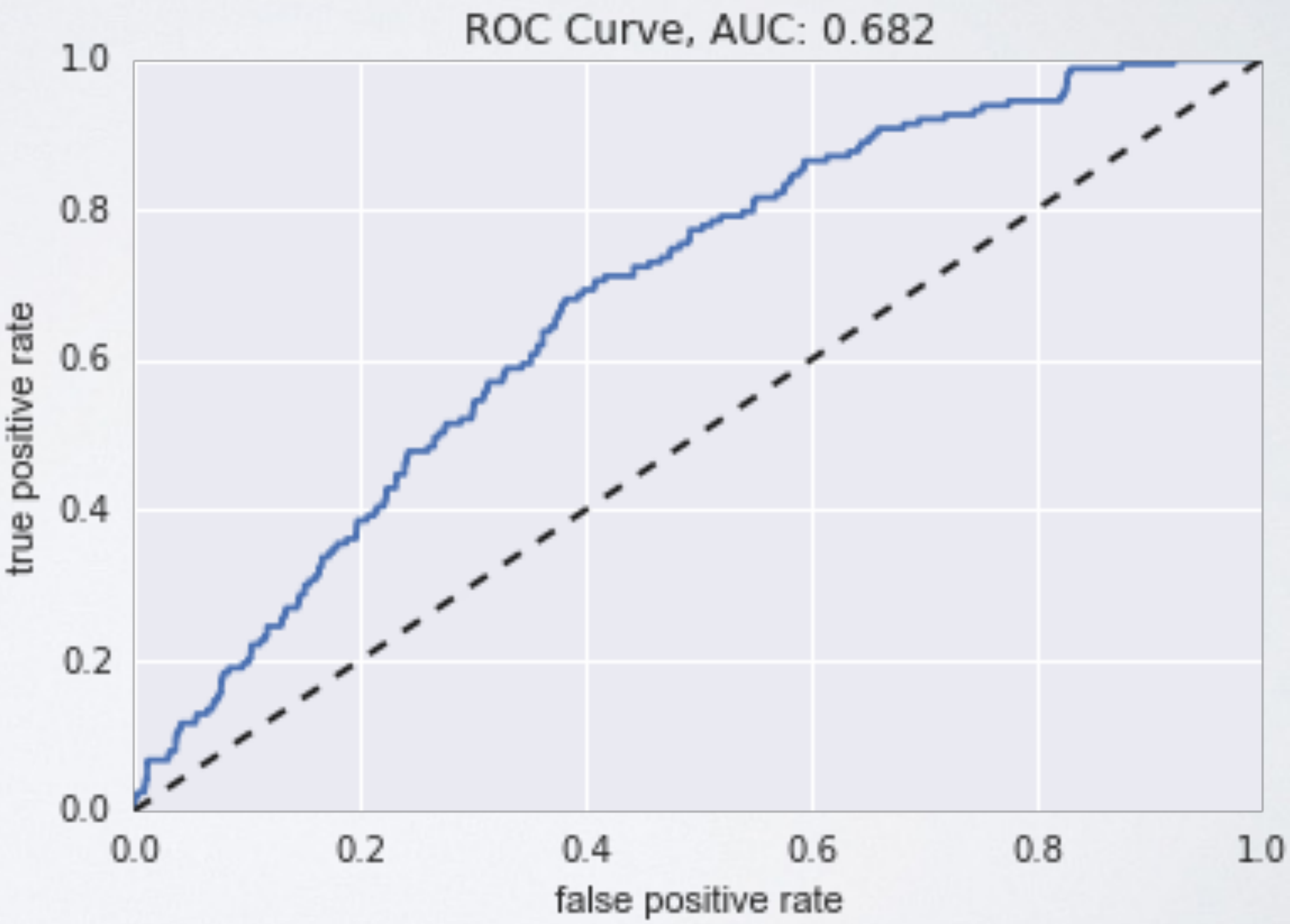
EXCLUDE ALL PHONE DATA

top decile esign rate: 0.123
average esign rate: 0.0649
top decile esign / avg. esign: 1.8958

F1 Score: 0.0435
accuracy: 0.9299
precision: 0.1905
recall: 0.0245

confusion matrix:

```
   Pred F  Pred T
F  [2332    17]
T  [ 159     4]
```



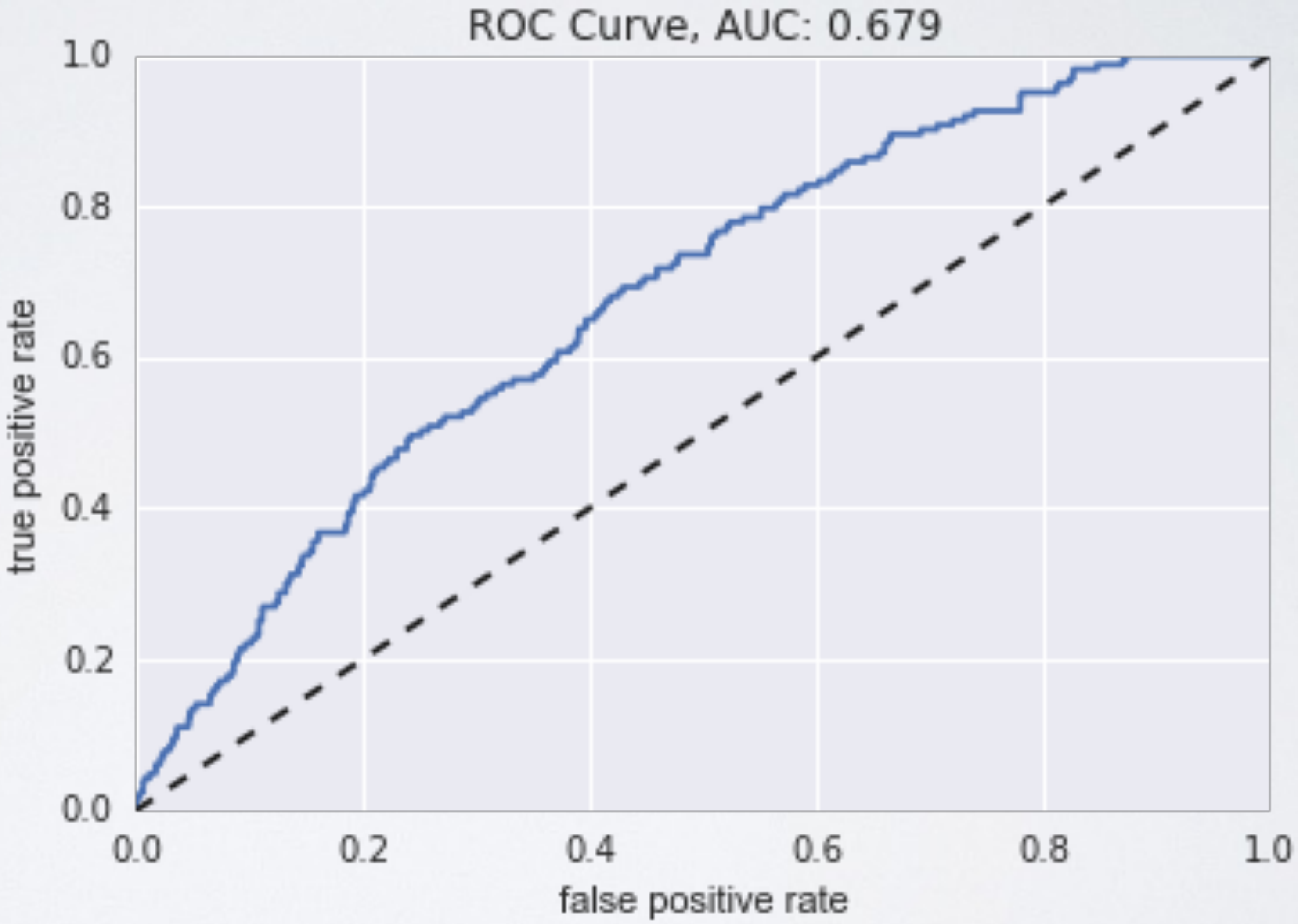
EXCLUDE ALL PHONE DATA, FEATURE IMPORTANCE > 0.001*

top decile esign rate: 0.1349
average esign rate: 0.0649
top decile esign / avg. esign: 2.1404

F1 Score: 0.0449
accuracy: 0.9323
precision: 0.2667
recall: 0.0245

confusion matrix:

```
   Pred F  Pred T
F   [2338    11]
T   [ 159     4]
```



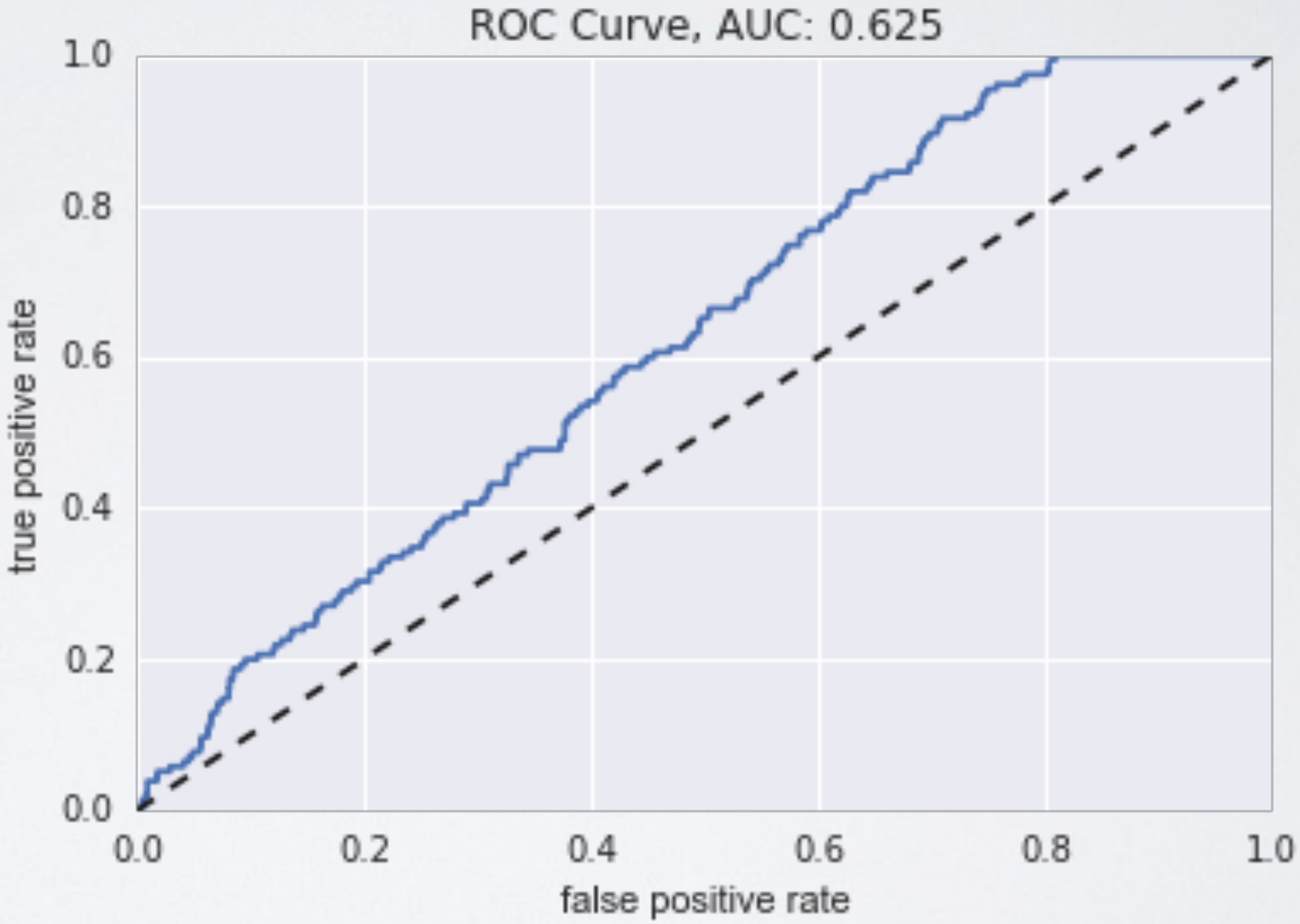
*173 to 121 features

EXCLUDE ALL ATTRIBUTION DATA

top decile esign rate: 0.0119
average esign rate: 0.0617
top decile esign / avg. esign: 1.9293

```
F1 Score:  0.0120
accuracy:  0.9347
precision:  0.0909
recall:     0.0065

confusion matrix:
   Pred F  Pred T
F   [2342   10]
T   [ 149    1]
```



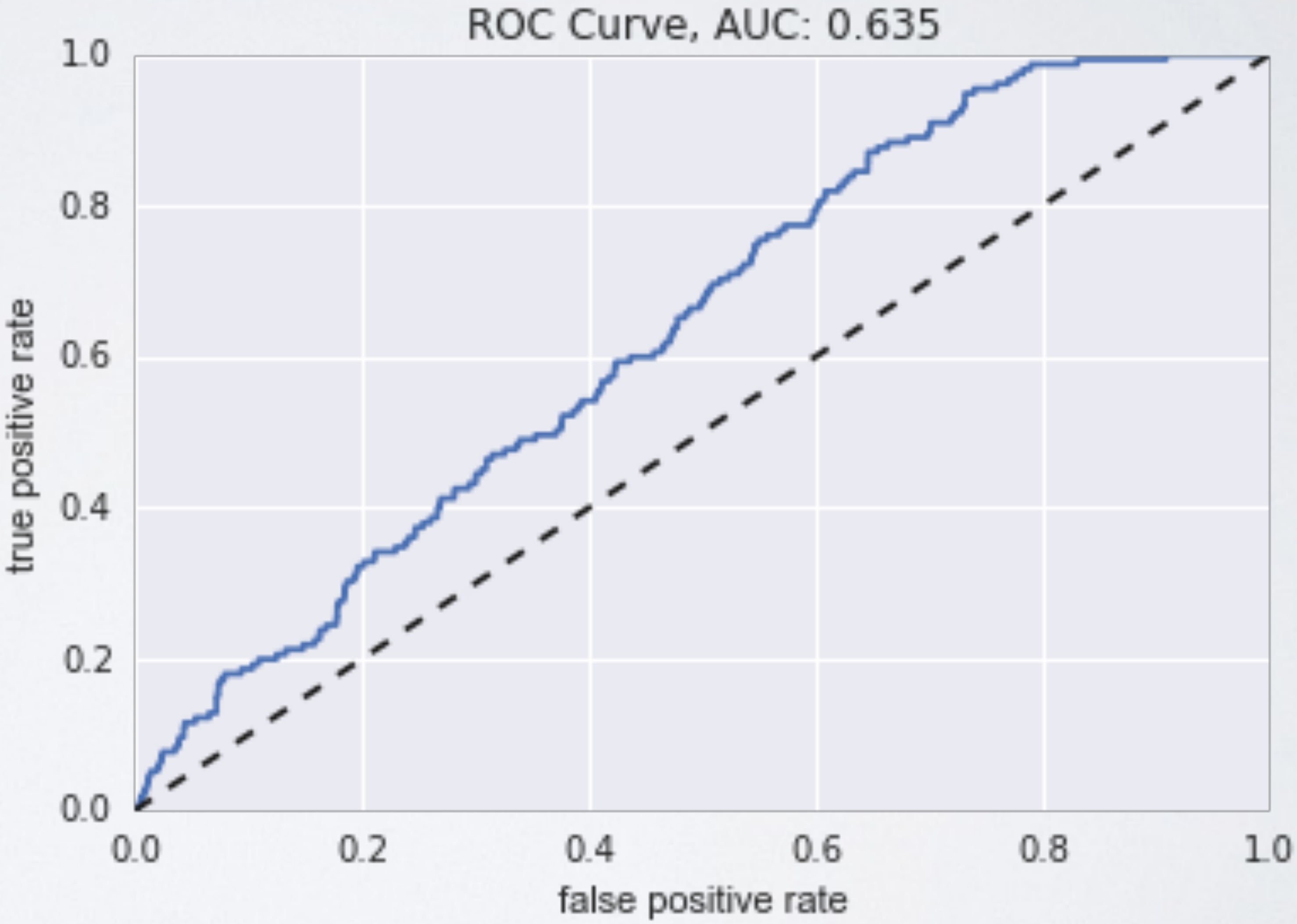
EXCLUDE ALL ATTRIBUTION DATA, FEATURE IMPORTANCE > 0.001*

top decile esign rate: 0.1151
average esign rate: 0.0617
top decile esign / avg. esign: 1.865

F1 Score: 0.0237
accuracy: 0.9343
precision: 0.1429
recall: 0.0129

confusion matrix:

$$\begin{array}{lcc} & \text{Pred F} & \text{Pred T} \\ \text{F} & [2345 & 12] \\ \text{T} & [\ 153 & 2] \end{array}$$



*141 to 98 features