```
@pavandantu18 →/workspaces/handson-10-machine-learning-with-mllib-pavandantu18 (main) $ python customer-churn-analysis.py
25/04/08 16:59:56 WARN Utils: Your hostname, codespaces-5dcf1c resolves to a loopback address: 127.0.0.1; using 10.0.11.29 instead (on interface eth0)
25/04/08 16:59:56 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/04/08 16:59:57 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
25/04/08 17:00:09 WARN InstanceBuilder: Failed to load implementation from:dev.ludovic.netlib.blas.JNIBLAS
25/04/08 17:00:09 WARN InstanceBuilder: Failed to load implementation from:dev.ludovic.netlib.blas.VectorBLAS
Logistic Regression AUC: 0.7137
Sample of selected features (top 5) and label:
+-------------------------+-----+
|selectedFeatures         |Churn|
+-------------------------+-----+
|(5,[1,4],[7.0,1.0])      |1    |
|(5,[1,2,3],[62.0,1.0,1.0])|0    |
|(5,[1,4],[5.0,1.0])      |0    |
|[1.0,6.0,1.0,1.0,0.0]    |1    |
|(5,[1],[61.0])           |0    |
+-------------------------+-----+
only showing top 5 rows
```

```
Starting cross-validation for LogisticRegression...
25/04/08 17:00:14 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC), users should configure it(them) to spark.eventLog.gcMetrics.youngGenerationGar
bageCollectors or spark.eventLog.gcMetrics.oldGenerationGarbageCollectors
LogisticRegression AUC: 0.7119
Starting cross-validation for DecisionTree...
DecisionTree AUC: 0.7628
Starting cross-validation for RandomForest...
RandomForest AUC: 0.7739
Starting cross-validation for GBT...
GBT AUC: 0.7786

Best Model Details:
Model: GBT
AUC: 0.7785714285714286
```

```
Best Hyperparameters: {Param(parent='GBTClassifier_f5cb510eb057', name='cacheNodeIds', doc='If false, the algorithm will pass trees to executors to match instances with nodes. If true, the algori
thm will cache node IDs for each instance. Caching can speed up training of deeper trees. Users can set how often should the cache be checkpointed or disable it by setting checkpointInterval.'):
False, Param(parent='GBTClassifier_f5cb510eb057', name='checkpointInterval', doc='set checkpoint interval (>= 1) or disable checkpoint (-1). E.g. 10 means that the cache will get checkpointed eve
ry 10 iterations. Note: this setting will be ignored if the checkpoint directory is not set in the SparkContext.'): 10, Param(parent='GBTClassifier_f5cb510eb057', name='featureSubsetStrategy', do
c="The number of features to consider for splits at each tree node. Supported options: 'auto' (choose automatically for task: If numTrees == 1, set to 'all'. If numTrees > 1 (forest), set to 'sqr
t' for classification and to 'onethird' for regression), 'all' (use all features), 'onethird' (use 1/3 of the features), 'sqrt' (use sqrt(number of features)), 'log2' (use log2(number of features
)), 'n' (when n is in the range (0, 1.0], use n * number of features. When n is in the range (1, number of features), use n features). default = 'auto'"): 'all', Param(parent='GBTClassifier_f5cb5
10eb057', name='featuresCol', doc='features column name.'): 'features', Param(parent='GBTClassifier_f5cb510eb057', name='impurity', doc='Criterion used for information gain calculation (case-inse
nsitive). Supported options: variance'): 'variance', Param(parent='GBTClassifier_f5cb510eb057', name='labelCol', doc='label column name.'): 'churn', Param(parent='GBTClassifier_f5cb510eb057', nam
e='leafCol', doc='Leaf indices column name. Predicted leaf index of each instance in each tree by preorder.'): '', Param(parent='GBTClassifier_f5cb510eb057', name='lossType', doc='Loss function w
hich GBT tries to minimize (case-insensitive). Supported options: logistic'): 'logistic', Param(parent='GBTClassifier_f5cb510eb057', name='maxBins', doc='Max number of bins for discretizing conti
nuous features.  Must be >=2 and >= number of categories for any categorical feature.'): 32, Param(parent='GBTClassifier_f5cb510eb057', name='maxDepth', doc='Maximum depth of the tree. (>= 0) E.g
., depth 0 means 1 leaf node; depth 1 means 1 internal node + 2 leaf nodes. Must be in range [0, 30].'): 3, Param(parent='GBTClassifier_f5cb510eb057', name='maxIter', doc='max number of iteration
s (>= 0).'): 10, Param(parent='GBTClassifier_f5cb510eb057', name='maxMemoryInMB', doc='Maximum memory in MB allocated to histogram aggregation. If too small, then 1 node will be split per iterati
on, and its aggregates may exceed this size.'): 256, Param(parent='GBTClassifier_f5cb510eb057', name='minInfoGain', doc='Minimum information gain for a split to be considered at a tree node.'): 0
.0, Param(parent='GBTClassifier_f5cb510eb057', name='minInstancesPerNode', doc='Minimum number of instances each child must have after split. If a split causes the left or right child to have few
er than minInstancesPerNode, the split will be discarded as invalid. Should be >= 1.'): 1, Param(parent='GBTClassifier_f5cb510eb057', name='minWeightFractionPerNode', doc='Minimum fraction of the
 weighted sample count that each child must have after split. If a split causes the fraction of the total weight in the left or right child to be less than minWeightFractionPerNode, the split wil
l be discarded as invalid. Should be in interval [0.0, 0.5).'): 0.0, Param(parent='GBTClassifier_f5cb510eb057', name='predictionCol', doc='prediction column name.'): 'prediction', Param(parent='G
BTClassifier_f5cb510eb057', name='probabilityCol', doc='Column name for predicted class conditional probabilities. Note: Not all models output well-calibrated probability estimates! These probabi
lities should be treated as confidences, not precise probabilities.'): 'probability', Param(parent='GBTClassifier_f5cb510eb057', name='rawPredictionCol', doc='raw prediction (a.k.a. confidence) c
olumn name.'): 'rawPrediction', Param(parent='GBTClassifier_f5cb510eb057', name='seed', doc='random seed.'): -1702011637598047055, Param(parent='GBTClassifier_f5cb510eb057', name='stepSize', doc=
'Step size (a.k.a. learning rate) in interval (0, 1] for shrinking the contribution of each estimator.'): 0.1, Param(parent='GBTClassifier_f5cb510eb057', name='subsamplingRate', doc='Fraction of
the training data used for learning each decision tree, in range (0, 1].'): 1.0, Param(parent='GBTClassifier_f5cb510eb057', name='validationTol', doc='Threshold for stopping early when fit with v
alidation is used. If the error rate on the validation input changes by less than the validationTol, then learning will stop early (before `maxIter`). This parameter is ignored when fit without v
alidation is used.'): 0.01}
```