Ques1. Regression Problem

Input features: 'age', 'sex', 'bmi','children', 'smoker', 'region'

Target feature: ‘charges’

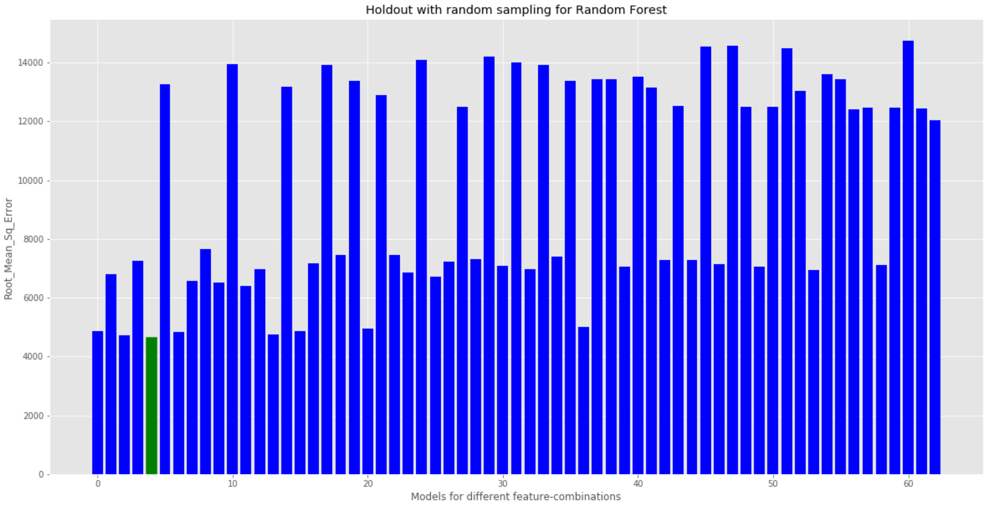
Regression Algorithm used:

- Random Forest
- XGBRegressor
- Linear Regression

Chosen error measure is RMSE, since error has the same dimensions as of the target output. Thus it’s easy to visualize the error received.

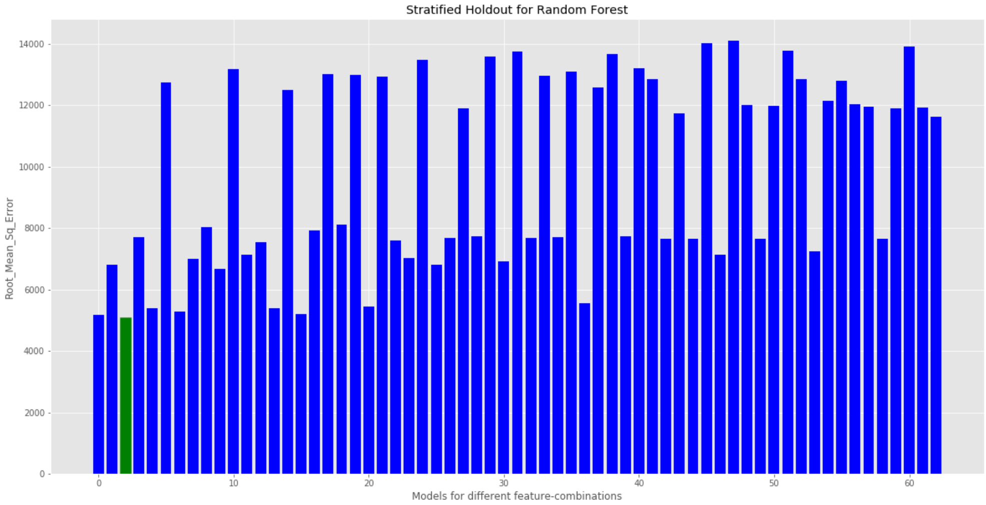
NOTE: The X-axis of the following plots corresponds to different combinations of input features, possible for first dataset. 6 features give 63 different combinations. X = 0 denote all features are considered and last value of X denote only one feature is considered. As value of X increases, size of the set decreases. Y-axis denotes the Root mean square error for a particular model.

1. Algorithm: Random Forest
   - Holdout method with random sampling:



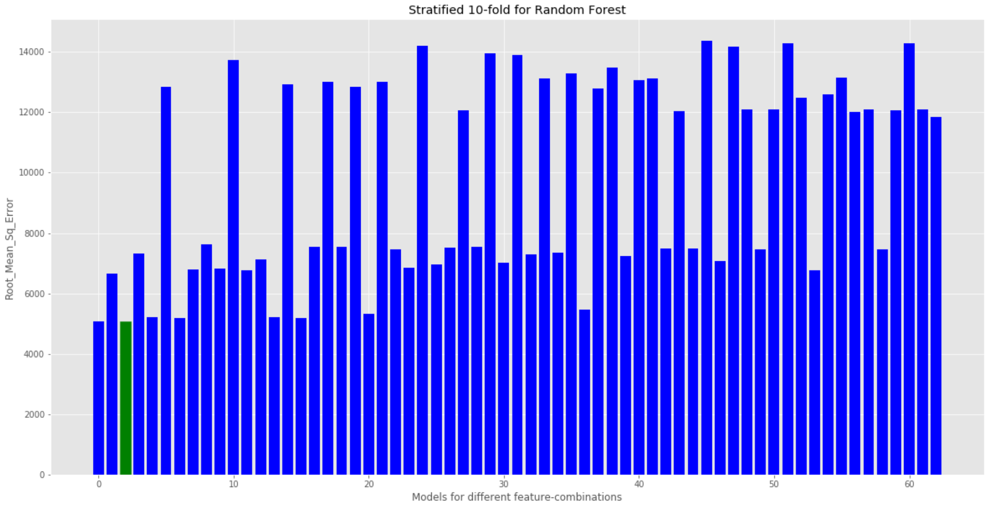
Min error value: 4688.8631

- Stratified holdout method:



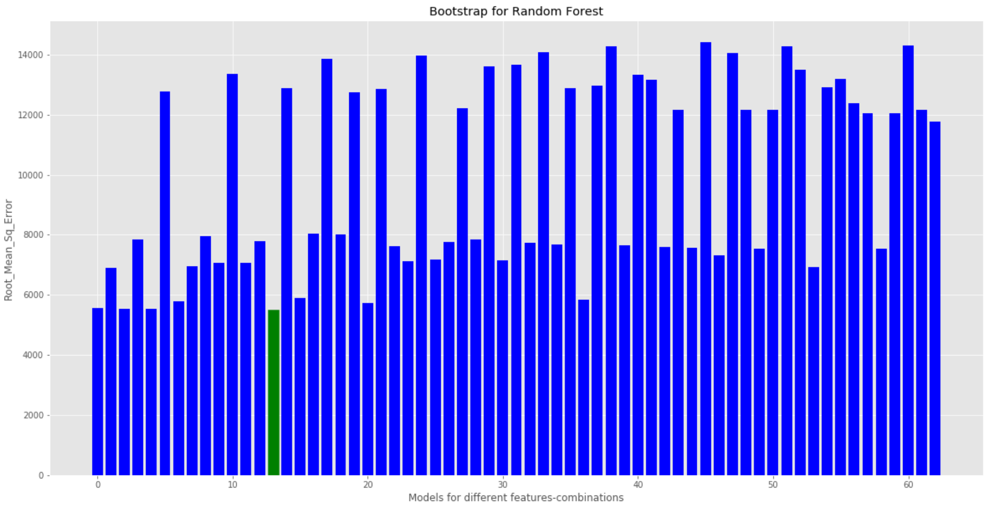
Min error value: 4921.3941

- Stratified 10-fold cross validation:



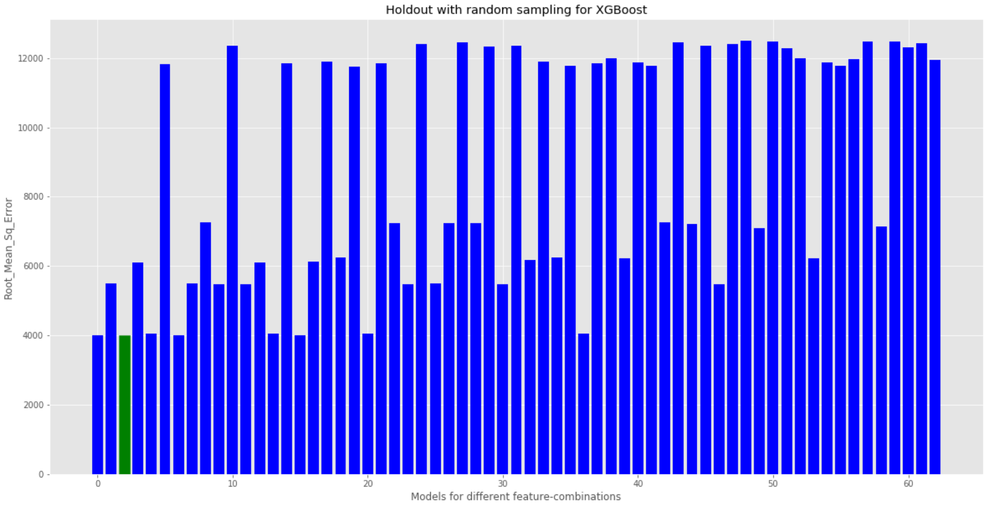
Min error value: 4993.7327

- Bootstrapping:



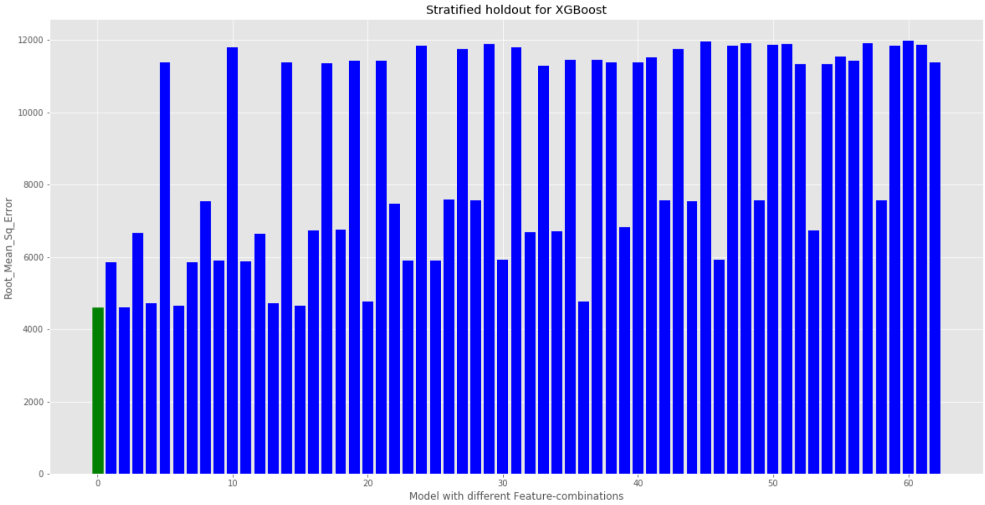
Min error value: 5509.1481

2. Algorithm: XGBRegressor
    - Holdout method with random sampling:



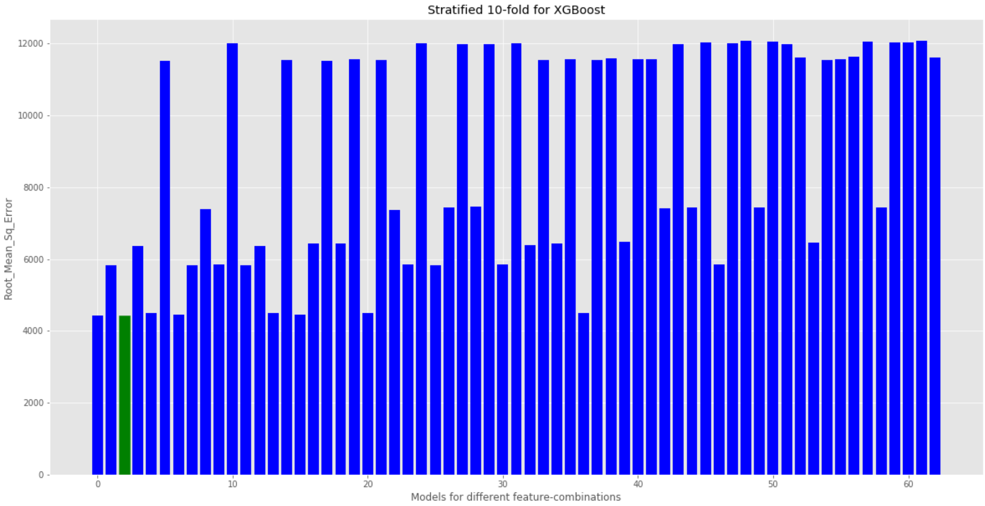
Min error value: 4004.4720

- Stratified holdout method:



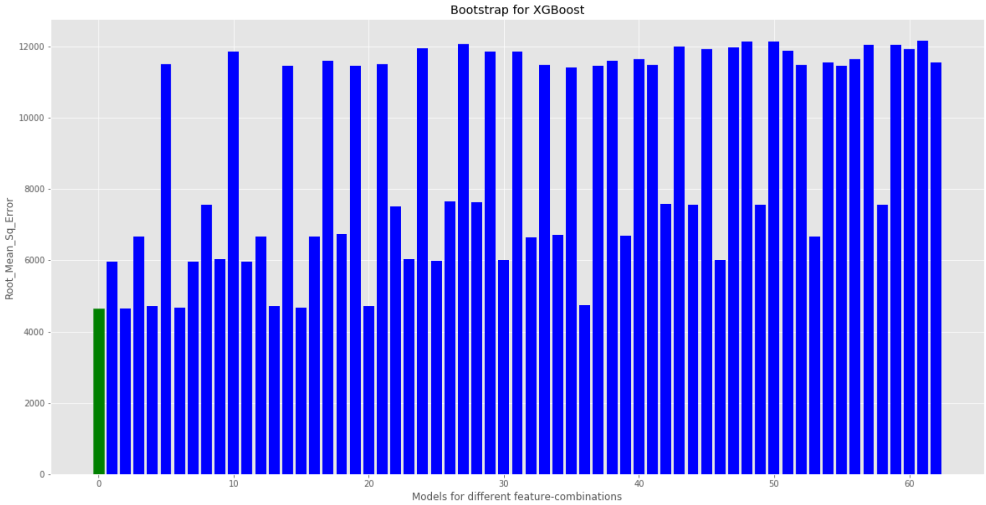
Min error value: 4603.7088

- Stratified 10-fold cross validation:



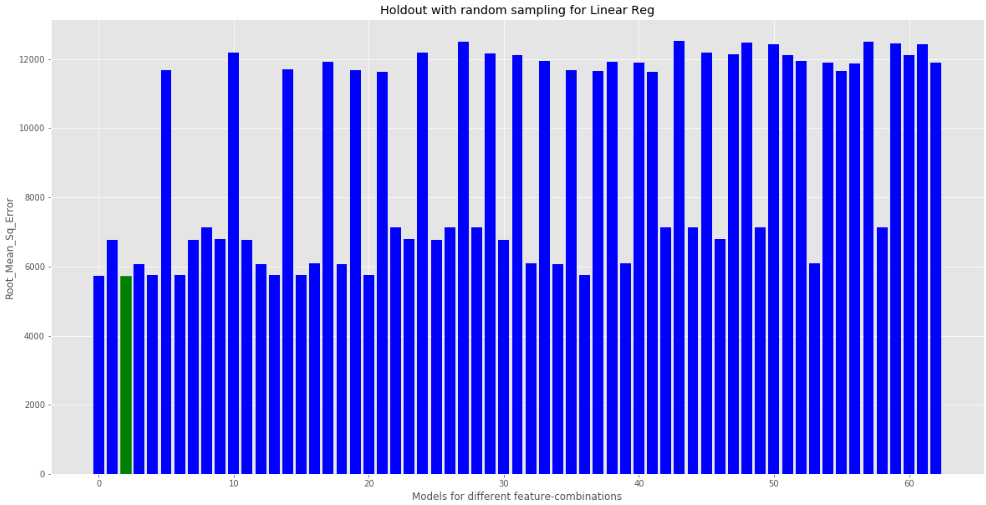
Min error value: 4428.8639

- Bootstrapping:



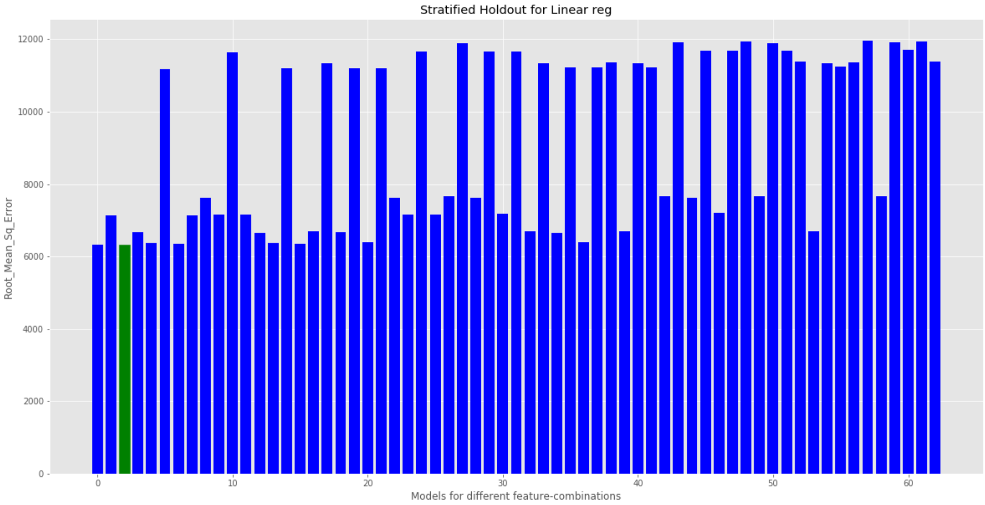
Min error value: 4652. 2818

3. Algorithm: Linear Regression
   - Holdout method with random sampling:



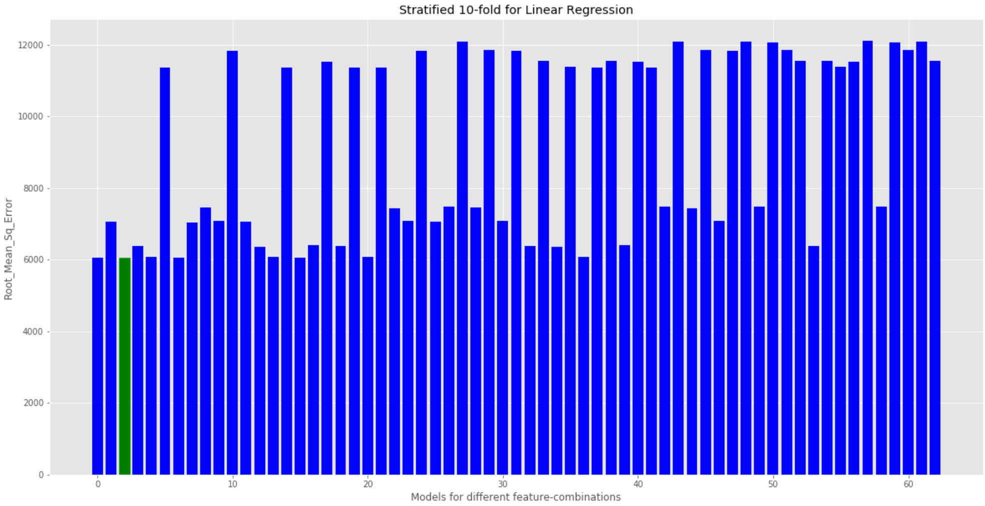
Min error value: 5731.6428

- Stratified holdout method:



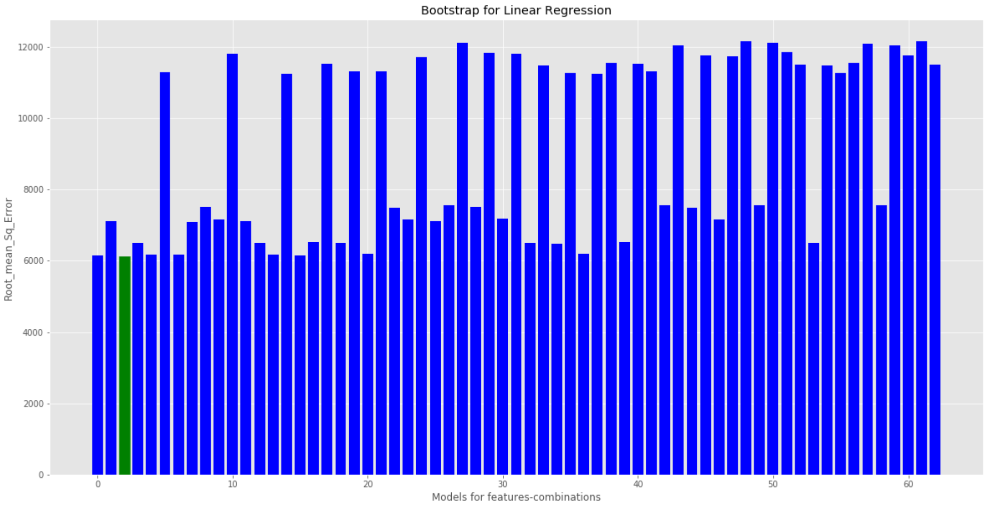
Min error value: 6330.3599

- Stratified 10-fold cross validation:



Min error value: 6049.6507

- Bootstrapping:



Min error value: 6139.9093

Out of all the models, minimum error is 4004.4720 , obtained for the following model:

1. Method: Holdout meathod with random sampling-
   - Test sample size is 40% of total data.
2. Regression algorithm is XGBRegressor:
   - Learning rate=0.01
   - n_estimators=1000
   - early_stopping_rounds=5
3. Input features: age, bmi, children, smoker, region

Ques 2. Classification Problem

Input features: 'MonthlyCharges', 'PaperlessBilling', 'OnlineBackup', 'TechSupport', 'OnlineSecurity', 'tenure', 'Contract'

Target feature: ‘Churn’

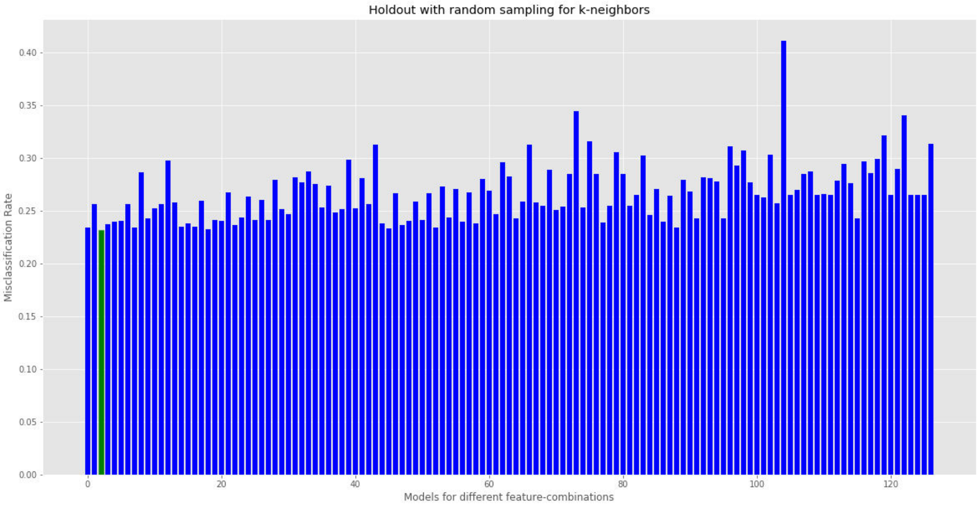
Regression Algorithm used:

- K-Neighbors Classifier
- Logistic Regression
- SVM Classifier

NOTE: Since number of input features in this dataset are 20, number of possible combinations exceeds computational limits of current machine. So 7 input features with high absolute correlation with target feature are chosen to be trained upon. They are stated above.

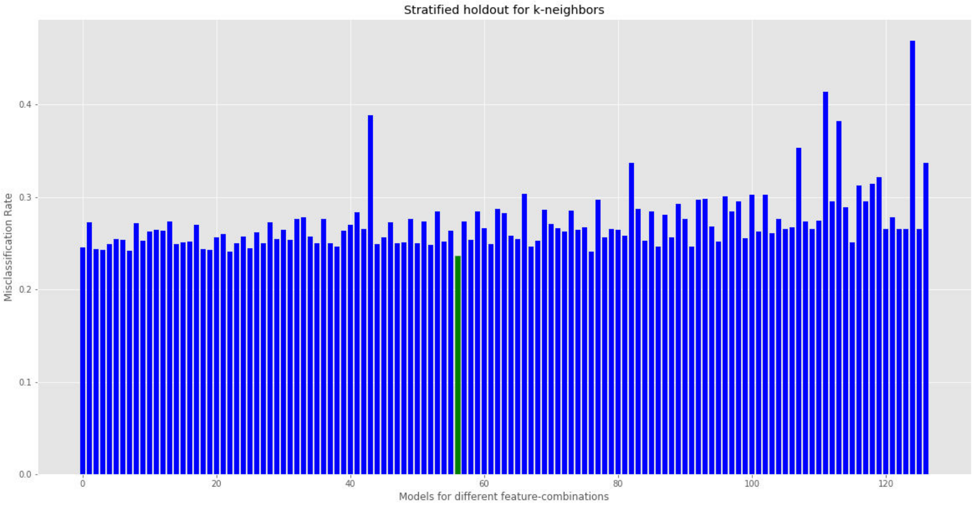
NOTE: The X-axis of the following plots corresponds to different combinations of input features, possible for first dataset. X = 0 denote all features are considered and last value of X denote only one feature is considered. As value of X increases, size of the set decreases. Y-axis denotes the misclassification rate for a particular model.

1. Algorithm: K-Neighbors Classifier
    - Holdout method with random sampling:



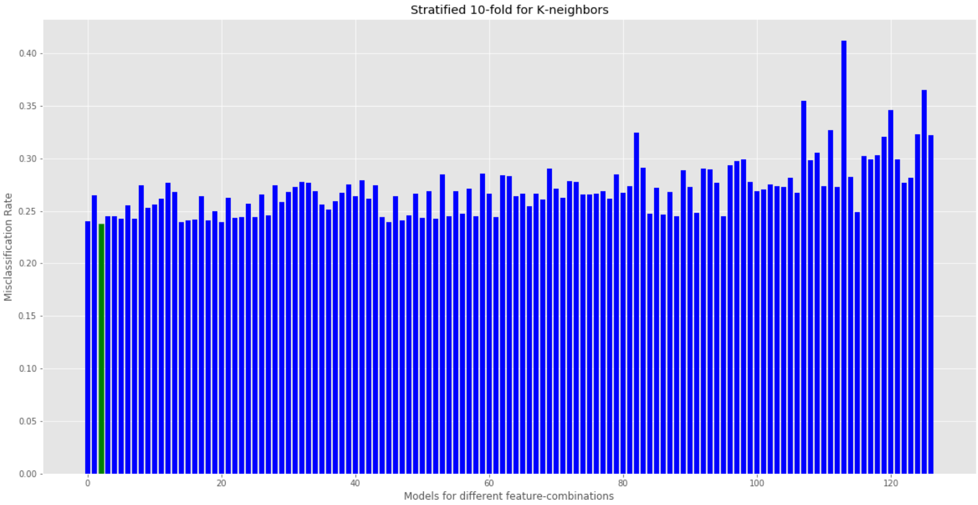
Min error value: 0.2317246712

- Stratified holdout method:



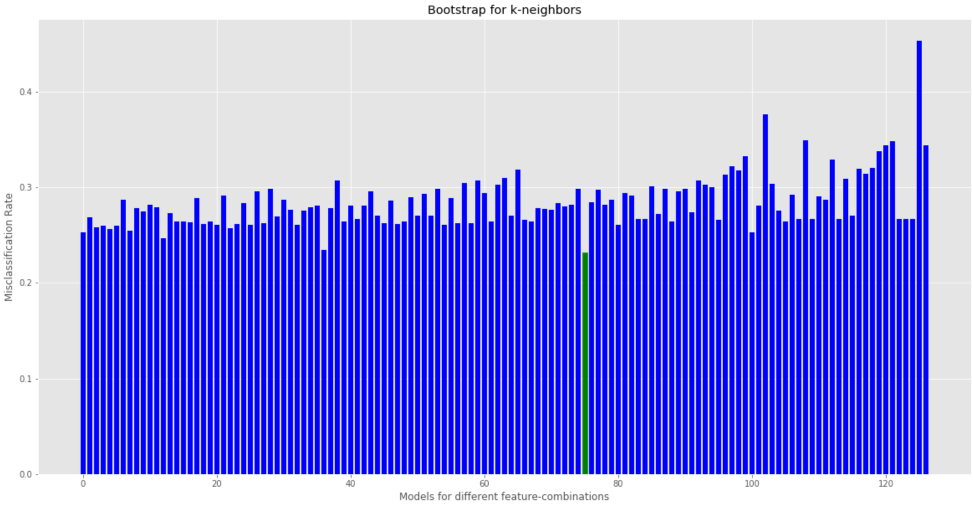
Min error value: 0.236337828

- Stratified 10-fold cross validation:



Min error value: 0.23782568

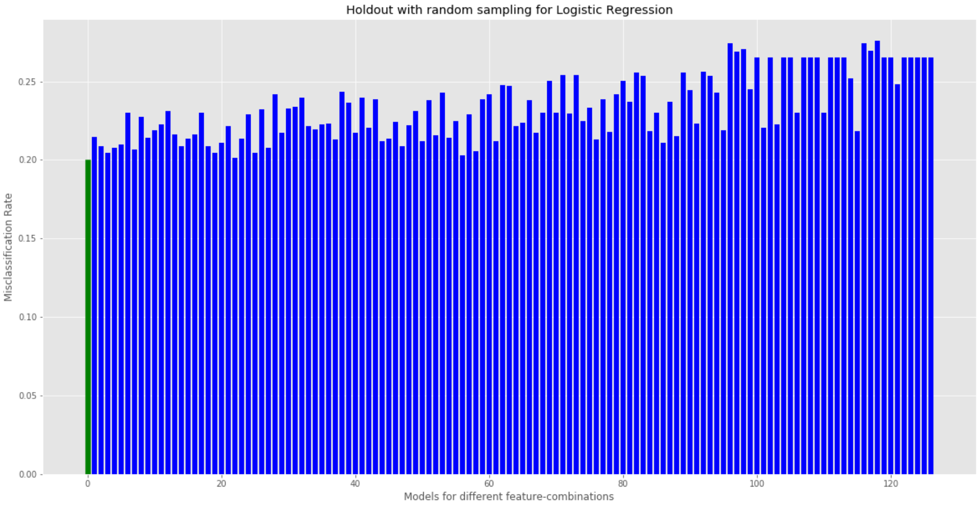
- Bootstrapping:



Min error value: 0.23143236

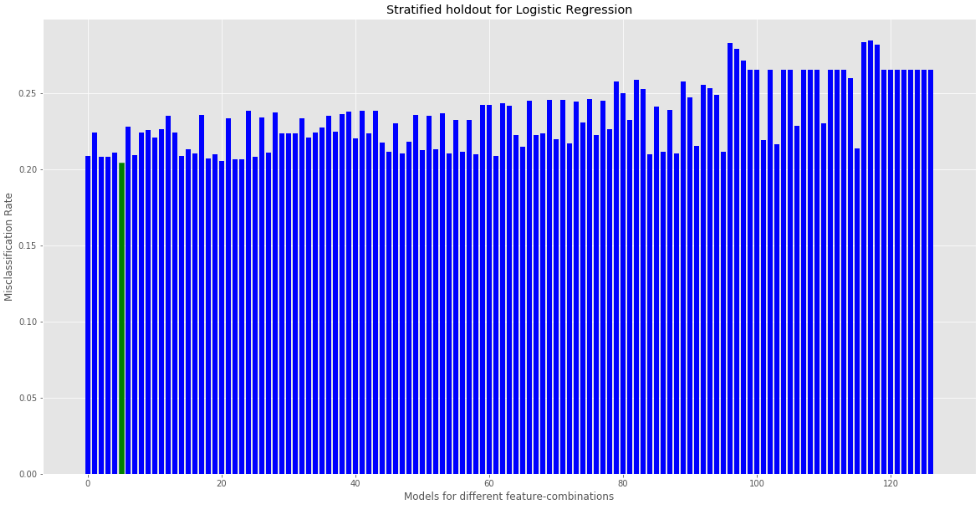
2. Algorithm: Logistic Regression

- Holdout method with random sampling:



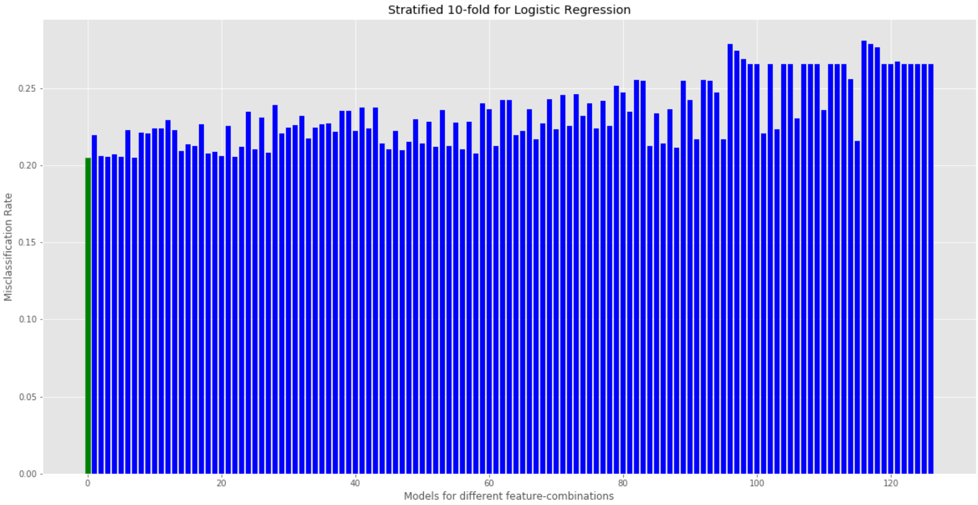
Min error value: 0.200496806

- Stratified holdout method:



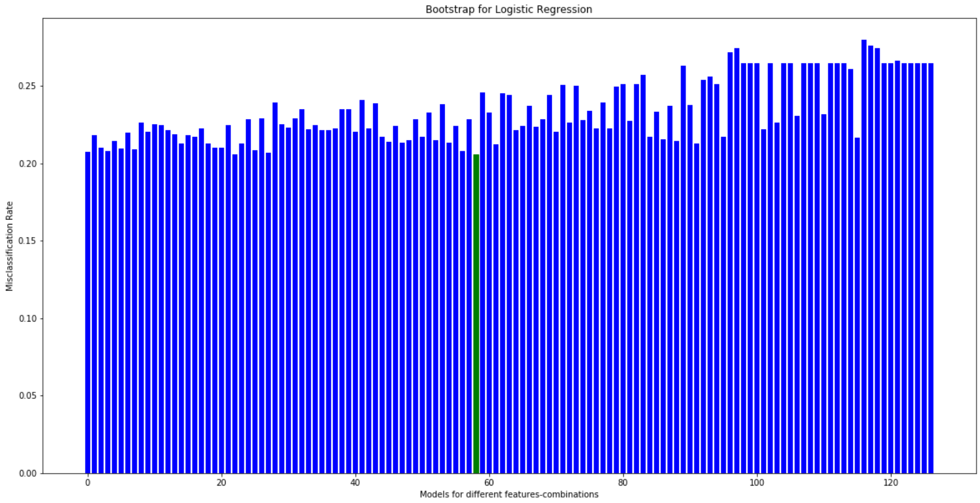
Min error value: 0.20440028

- Stratified 10-fold cross validation:



Min error value: 0.20474258

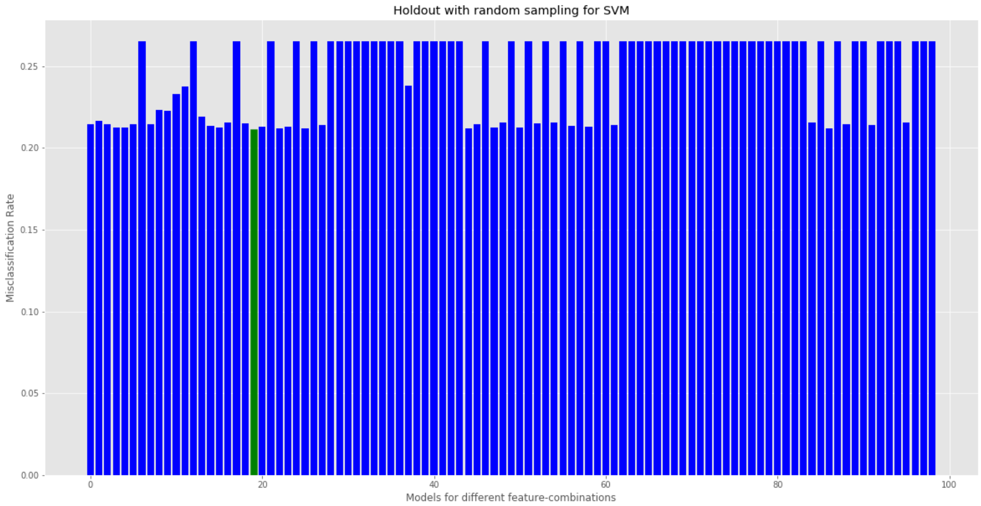
- Bootstrapping:



Min error value: 0.20565053

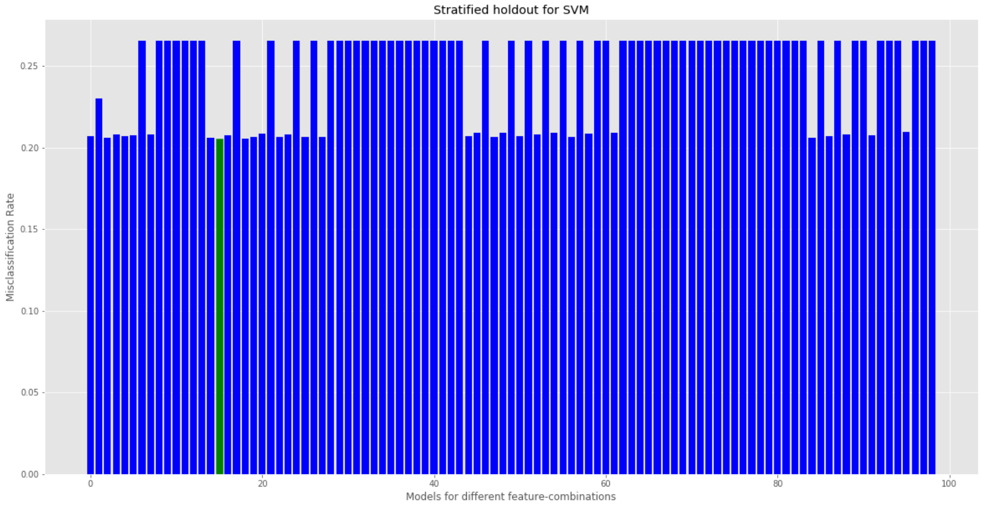
3. Algorithm: SVM Classifier

- Holdout method with random sampling:



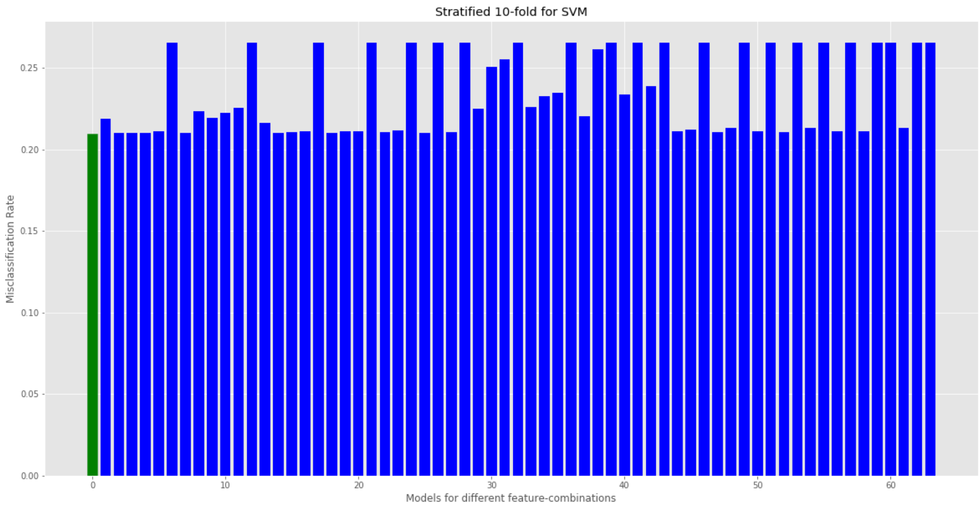
Min error value: 0.21149751

- Stratified holdout method:



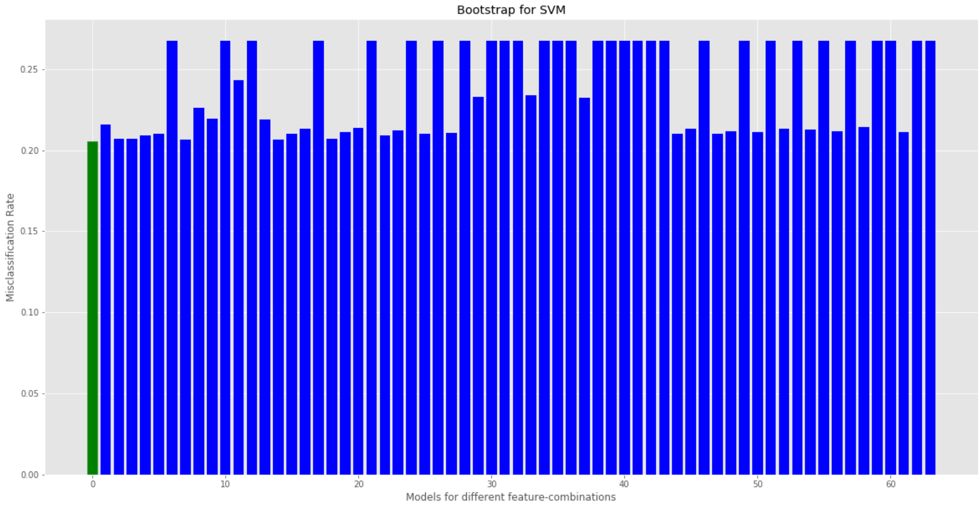
Min error value: 0.20581973

- Stratified 10-fold cross validation:



Min error value: 0.20971256

- Bootstrapping:



Min error value: 0.20557029

Out of all the models, minimum misclassification rate is 0.20440028, observed for the following model:

1. Method: Stratified holdout method-
2. Test sample size is 40% of total data.
3. Stratified wrt ‘Churn’(target feature)
4. Regression algorithm is Logistic Regression.
5. Input features: 'MonthlyCharges', 'PaperlessBilling', 'OnlineBackup', 'TechSupport', 'tenure', 'Contract'