Hackathon Report

This is a brief report describing the work I have done.

Initially I developed a basic outline model by converting all categorical features to numerical (Removing highly out of the way features). I got an accuracy of 81.33 at this stage.

Then I have analysed data carefully and further removed columns by eyeballing each and every feature. I tried to keep features as many low as possible which is practically done in industry (Talked to my friend who is working in Financial Data analytics.)

Steps followed in Feature Engineering and data cleaning :

- Implemented Information Value Weight of Confidence
- Created New Features using the Binned method. Tried out with different combinations
- Tried with Mode and Mean (To fill null values)
- Used Scatter plots to check out outliers
- Used frequency of distribution to check out about features
- Tried Dropping columns with correlation values
- Implemented Chi-square method in Feature Selection
- Used Feature Importance matrix to remove columns

I initially went with Naive Bayes (as it is suitable for a data set with multiple features) Then I went with SVM (as there are a lot of features in dataset, SVM suits) and Logistic regression.
But with correlation matrix and other means I understood all features are weakly affecting the selection ,So went with Ensemble methods.

Then worked with Gradient Boosting and Random Forests. Random Forests started showing a bit of higher accuracy compared to Gradient Boosting. Used both Cross validation, train_test_split method to take care of overfitting while building models. Built even other classification models to have a better understanding. Finally achieved an accuracy of **86.057 .**

Unfortunately there are no standardised methods at this stage and everything is trial and error based, I have tried all the above mentioned methods in different combinations of which few worked out, few didn't. But a great practical learning experience.

Kaggle Id: 8856256
SM21MTECH12012
H.N Srikanth

Proof that I worked with different techniques.