Hire Analytics

Problem Statement:

Human Resources are critical resources of any organization. Organizations spend a huge amount of time and money to hire their employees. It is a huge loss for companies if the company reject the right candidate or hire the wrong candidate, so if the HR can predict whether to hire a candidate or not is an optimal way, it will reduce the time and cost to hire a candidate.

Objective:

The objective of the present machine learning model is to study various factors of candidates by which it will help the **HR** know the right candidate for organization and factors associated with their selection.

General methodology

This problem comes under the binary classification . Please consider the following points and assumptions while solving the problem:-

- The data is very small in size .
- The detail of all feature is not available so we consider all categorical variable as a nominal category.
- There is no null value in data but the "?" value comes in multiple columns so we consider this value as null and use mode for a categorical feature and median for the numerical feature.
- There is a lot of an outlier in numerical columns(C15 have most) but we keep outlier because data is very less and we don't know about all the actual columns which represent as C1, C2…..etc.
- We use multiple models for the problem but Random Forest Classifier gives the best result with fine-tuning.
- We use the F1 score as a scoring matrix.

Justification for selected variables

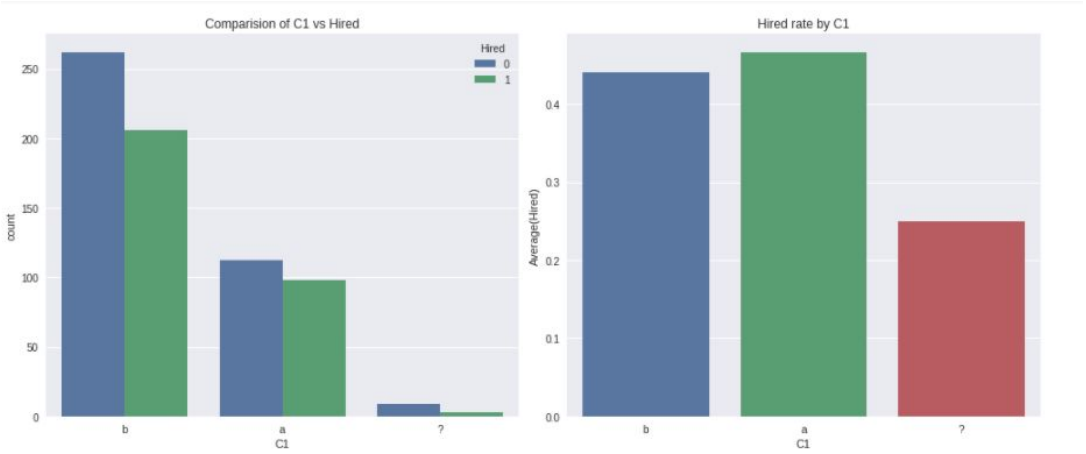
We have use ANOVA test for numerical feature and Chi2 test for categorical feature to select the best features in data and we also get very low score for numerical column (C14) and categorical columns(C1,C7,C12),so we drop these 4 columns from the final data set.

For selected column visuals, refer to code..

Please find the EDA for these four dropped columns in coming slides.

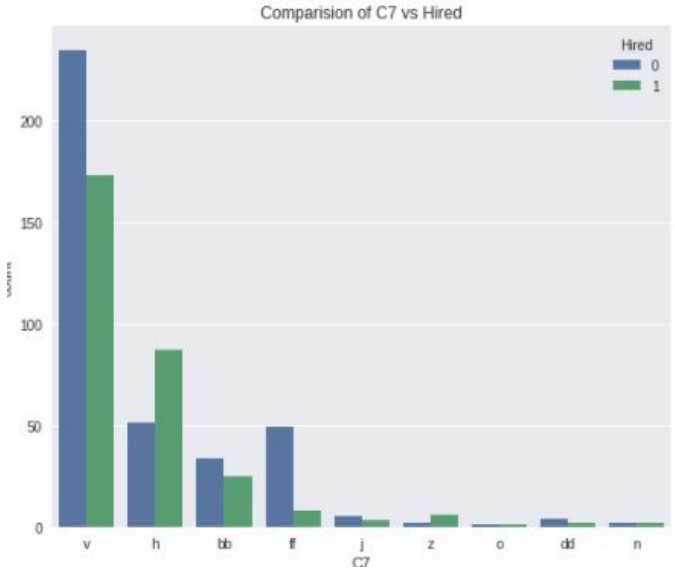
Feature C1

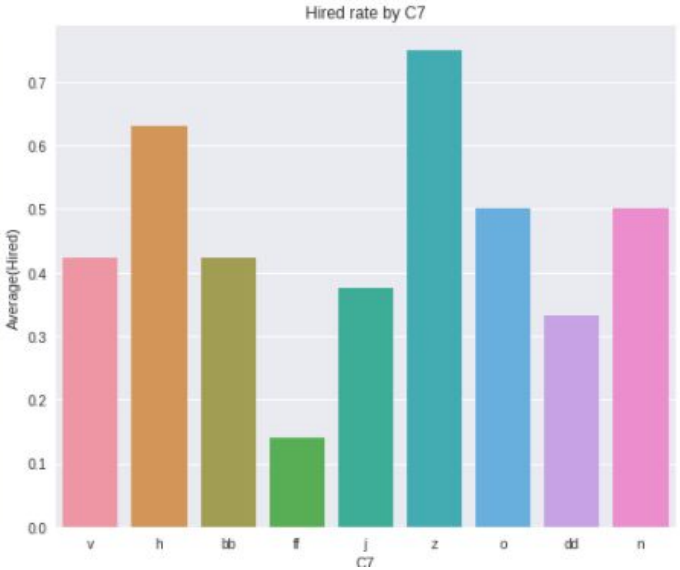
Hire rate for C1 is same for ‘a’ and ‘b’ value as show in below chart.As C1 has no significant effect on hiring,so it will not be taken under consideration.



Feature C7

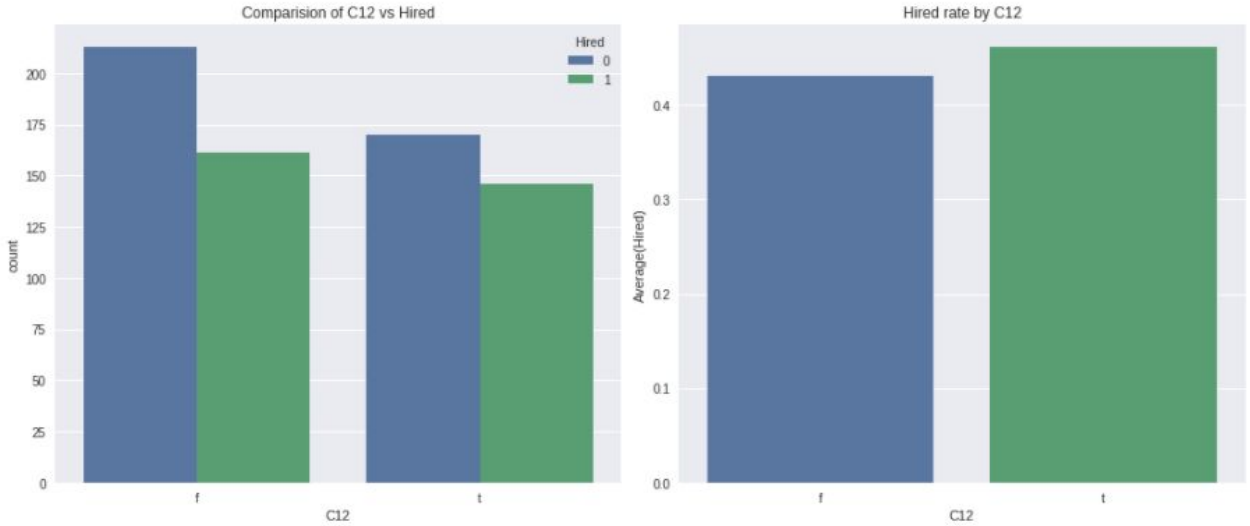
Hire rate for C7 doesn’t have any pattern as shown in below chart. As C7 has no significant effect on hiring, so it will not be taken under consideration.





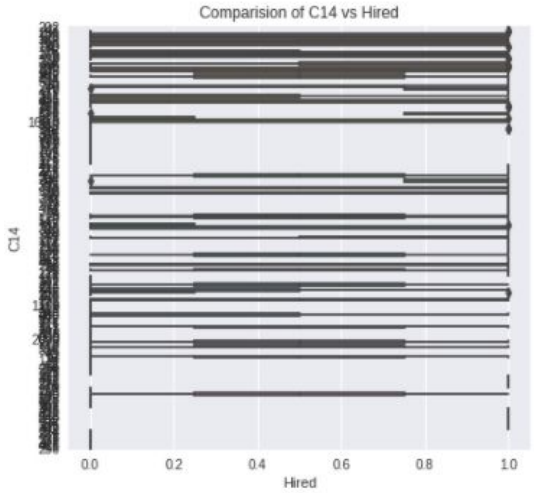
Feature C12

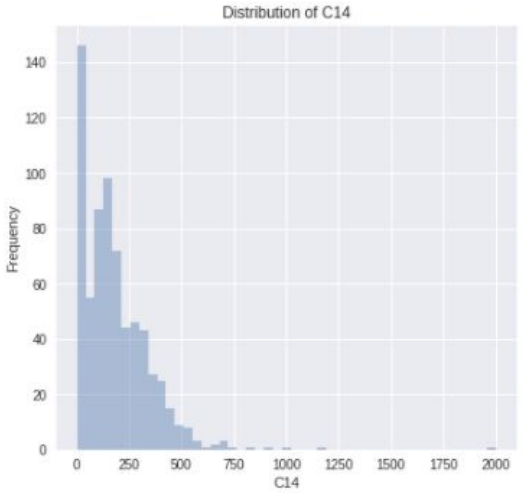
Hire rate for C12 doesn’t have any pattern as shown in below chart. As C12 has no significant effect on hiring,so it will not be taken under consideration.



Feature C14

Distribution of C14 is more towards zero and by using ANOVA test C14 shows very less impact on model, so we dropped C14 in final data set.





Justification for accuracy measure

As we mention in the Problem statement hiring the wrong candidate or losing the right one both cost the company . So that is why False-negative rate and False-positive rate are crucial for this problem .Considering this we use **F1-score** as an accuracy measure .

	precision	recall	f1-score	support
Non hired	0.84	0.90	0.87	68
Hired	0.89	0.83	0.86	70
accuracy			0.86	138
macro avg	0.86	0.86	0.86	138
weighted avg	0.86	0.86	0.86	138

Justification for algorithm(s)

For given data we use below algorithm:

- LogisticRegression
- GradientBoostingClassifier
- KNeighborsClassifier
- RandomForestClassifier
- XGBClassifier

We get best result using Random Forest Classifier and used RandomizedSearchCV to fine tune the model .