STATE FARM INTEREST RATE PREDICTIONS

Vybhav reddy kc

FLOW OF PRESENTATION

- Environment Used
- Data Description
- Preliminary Analysis
- Model & Evaluation
- Future Improvements and suggestions
- Python code and execution
- References

ENVIRONMENT USED

- Python- 2.7.11
  - Packages : Pandas, Seaborn, Numpy, Sklearn ,Ref.estimators

- Anaconda – Spyder 2.8 development platform
- Jupytor – Ipython 4.0.3

DATA DESCRIPTION

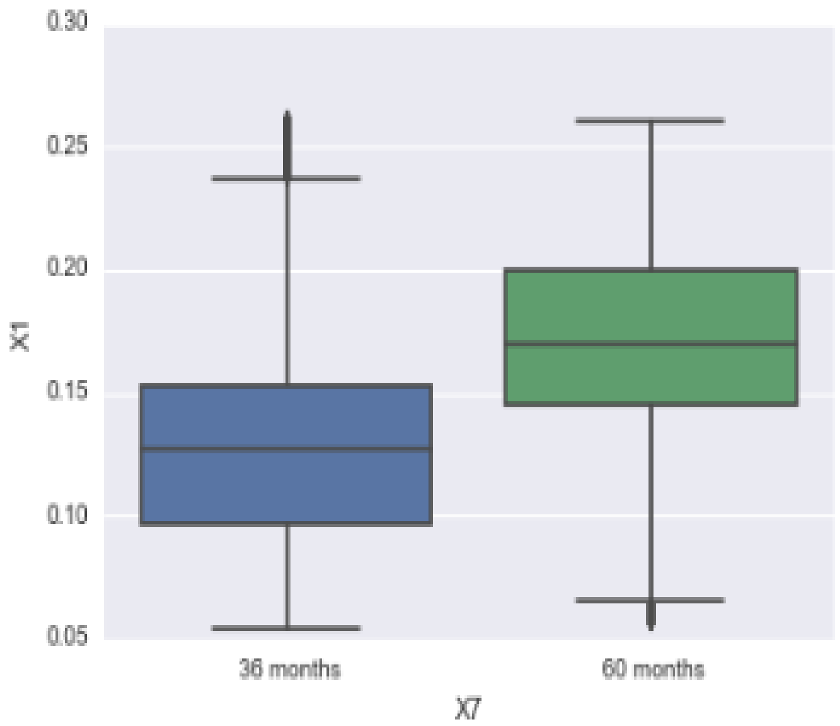
- Training data 400,000 records
- Testing data 80,000 records
- 31 features which contain loan amount requested, amount funded, investor funded amount, Number of payments, Loan grade, Loan subgrade, Reason for loan, loan category, Loan title, zipcode, date loan issued, state and other features

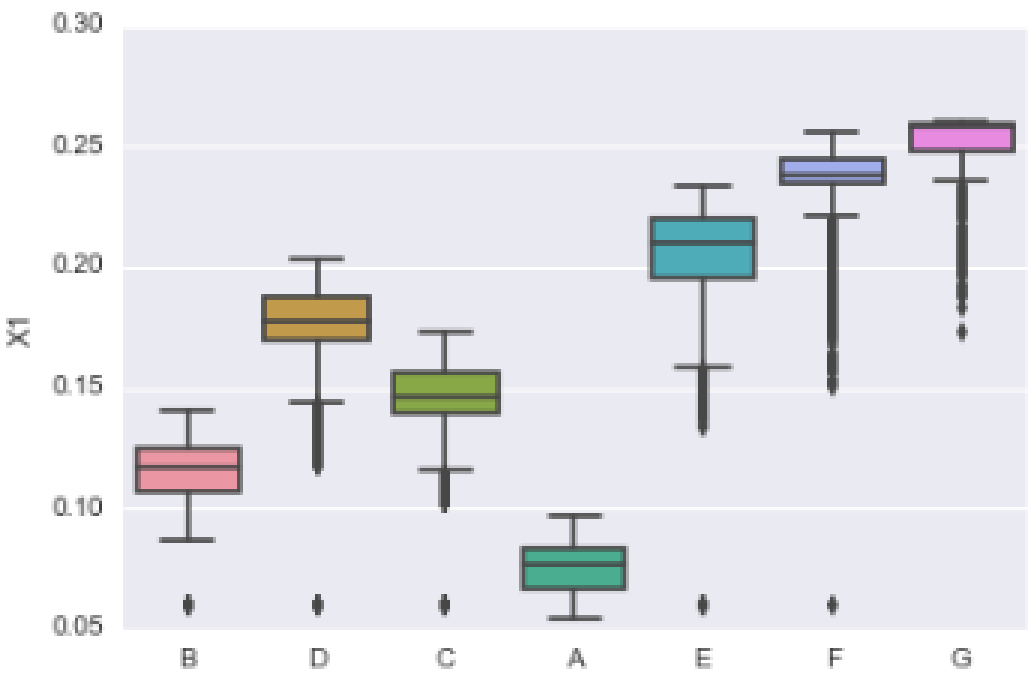
- Target – Interest rate of the loan
- Categorical features - 13 features
- 48 dummy features are created for categorical features.

PRELIMINARY ANALYSIS

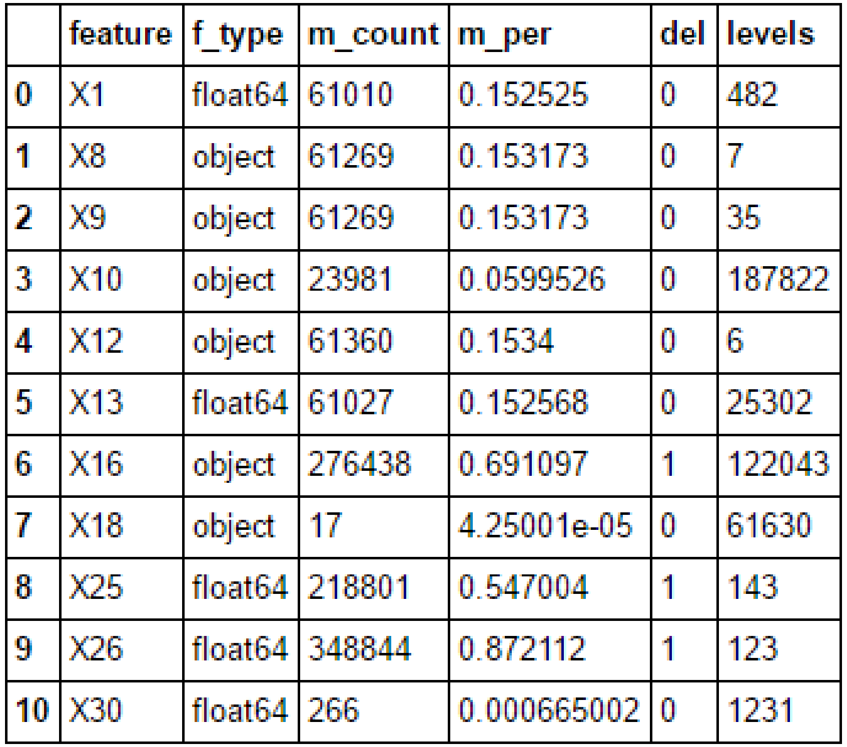
- Removing the special characters
- ‘$’,’%’, ‘,’
- Years of Experience
- ‘<1 year’,’10+ years’, ‘n/a’

OUTLIERS


0.30
0.25
0.20
0.15
0.10
0.05
X1
36 months
60 months
X7


0.30
0.25
0.20
0.15
0.10
0.05
X1
B
D
C
A
E
F
G

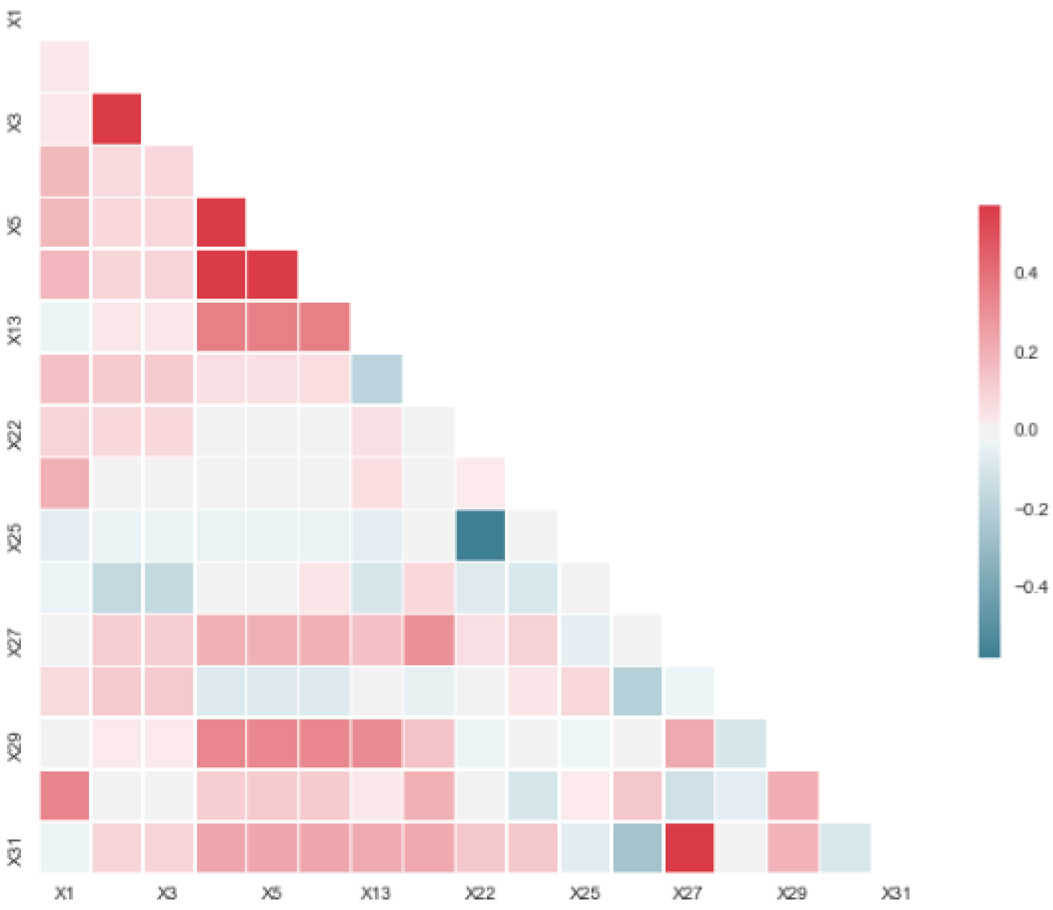
MISSING VALUES



	feature	f_type	m_count	m_per	del	levels
0	X1	float64	61010	0.152525	0	482
1	X8	object	61269	0.153173	0	7
2	X9	object	61269	0.153173	0	35
3	X10	object	23981	0.0599526	0	187822
4	X12	object	61360	0.1534	0	6
5	X13	float64	61027	0.152568	0	25302
6	X16	object	276438	0.691097	1	122043
7	X18	object	17	4.25001e-05	0	61630
8	X25	float64	218801	0.547004	1	143
9	X26	float64	348844	0.872112	1	123
10	X30	float64	266	0.000665002	0	1231

- Loan grade , annual income, Home owner status are closely related and doesn’t have a specific pattern.
- Delete the columns which has more than 50 % of missing values
- Replacing the missing values in categorical features with most repeated values
- Replacing the missing values in continuous features with median values.

CORRELATION BETWEEN FEATURES



- Features X4 , X5, X6 are high correlated values

MODELLING:

- After the pre-processing steps, dataset has 74 columns and 338989 rows.
- Divided the data into train and validation with 75:25 ratio
- Algorithms that are resistant to missing and outlying values have to be chosen.
- Under and over fitting of algorithms can be controlled using cross validation.

MODEL EVALUATION

- Metric : Root Mean Square Error

$$\text{RMS} = \sqrt{\frac{\sum\left(X_{\text{Observed}} - X_{\text{Grid}}\right)^2}{n\,(\text{number of samples represented by grid})}}$$

- **Ridge Regression with 10 fold CV**
  - Train : 0.826672
  - Validation : 0.831106
  - RMSE : 0.018050

- **RandomForest Regressor**
  - With 10-fold Cross validation
  - RMSE : 0.012511

StateFarm®

Preventing crashes, reducing injuries and saving lives

FUTURE IMPROVEMENTS AND SUGGESTIONS

- Due to time constraints feature engineering could not be done, engineering features could be of greater value to the model.
- Missing values can be imputed (using knn, random forest or SVM) better instead of imputing with mean and median.
- Multiple models can be run instead of just two.
- Random forest bag sizes can be tuned to improve the number of data reads, given more time.

PYTHON CODE AND EXECUTION

- SourceCode:
- main( )
- data_preprocessing( )
- data_modeling( )
- RMSE( )

- main function will load the data and processes it, saves the predictions to CSV file.
- Note:
- The Data for cleaning & modeling, Holdout data for testing has to be in the same folder
- where the source code is placed.
- To run the code, use the below in python editor
- python SourceCode.py

REFERENCES

- http://pandas.pydata.org/pandas-docs/stable/index.html
- https://web.stanford.edu/~mwaskom/software/seaborn/
- http://scikit-learn.org/stable/index.html
- https://github.com/rasbt/python-machine-learning-book