GUSTAVO DAMASCO / DS.SF.39

FINAL PROJECT PRESENTATION

SAN FRANCISCO / 12.05.17

FINAL PROJECT

YOUNG PEOPLE SPENDING HABITS

PROJECT OVERVIEW

Can we **predict spending habits** of a person from his/hers **personality traits, views on life & opinions**?

Applications:

- Helpful to understand what behaviors / characteristics of someone influenced their spending habits.
- For companies, could be helpful to target customers as good or bad money savers.

DATA

- Data is from a survey made with 1010 students in 2013.
- All participants were from Slovakian nationality, aged between 15-30.
- The variables can be split into the following groups:
  - (1) **music preferences**, (2) **movie preferences**, (3) **hobbies & interests**, (4) **phobias**, (5) **health habits**, (6) **personality traits, views on life & opinions**, (7) **spending habits** and (8) **demographics**.

EXPLORING THE DATA

- Loaded the full dataset into Pandas

```
df.shape
```

```
(1010, 150)
```

- Selected the columns from the **personality traits, views on life & opinions** and **spending habits** groups.

```
df_spending.shape
```

```
(1010,)
```

```
df_personality.shape
```

```
(1010, 57)
```

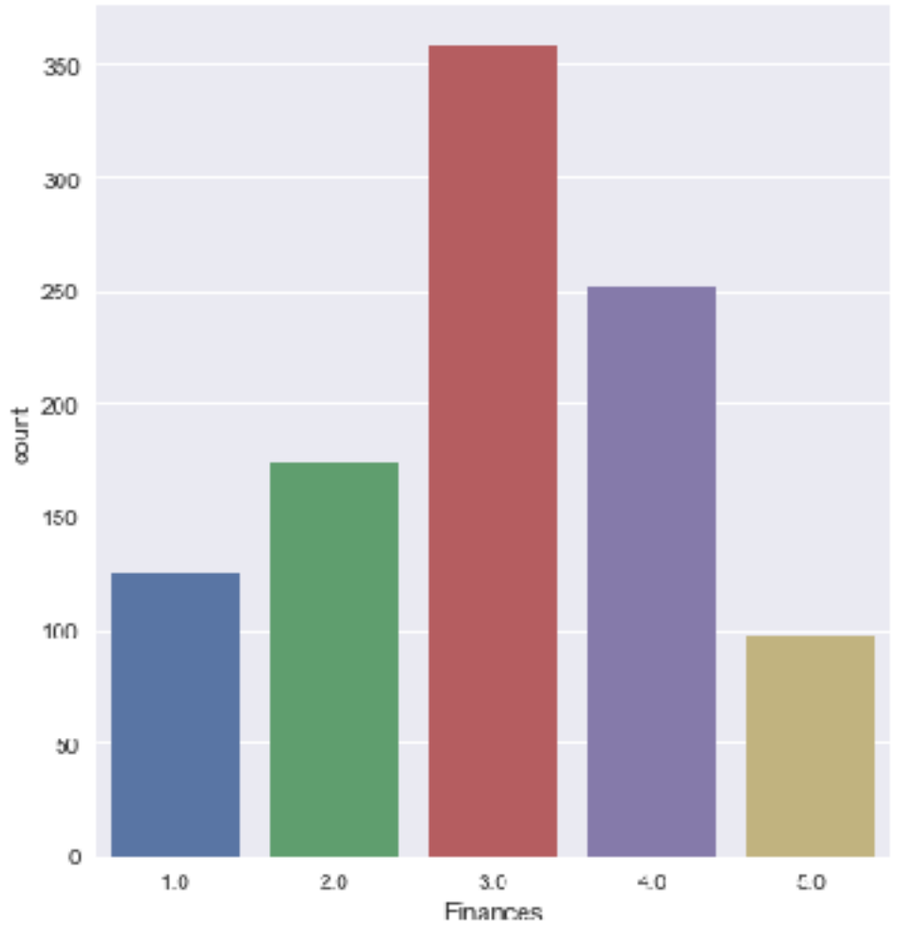
- Created a dataset with only the columns from those **two groups**.

```
df_data.shape
```

```
(1010, 58)
```

EXPLORING THE DATA

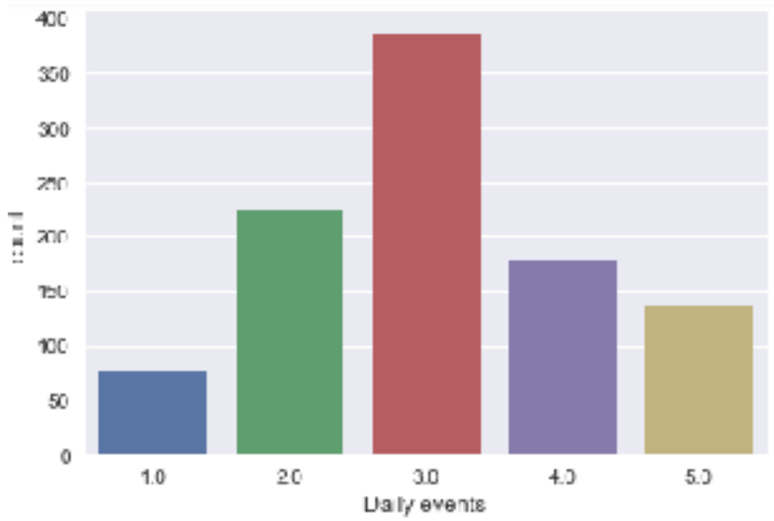
- Dependent variable distribution:
  - I save all the money I can (Strongly disagree 1-2-3-4-5 Strongly agree)

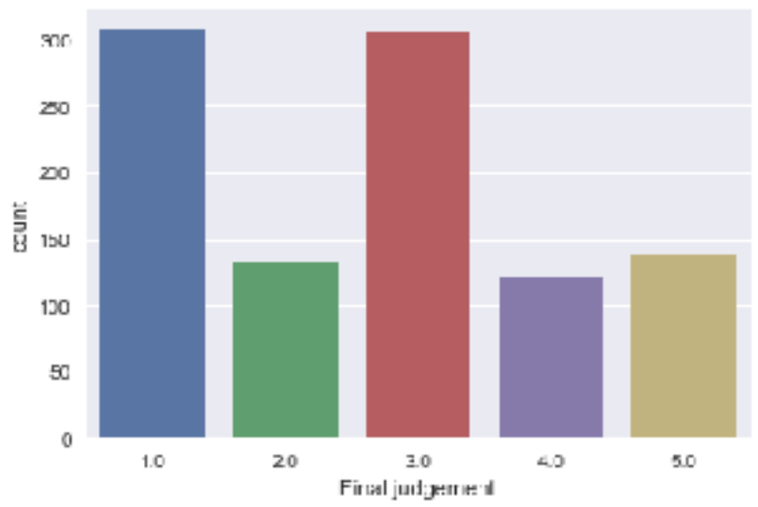


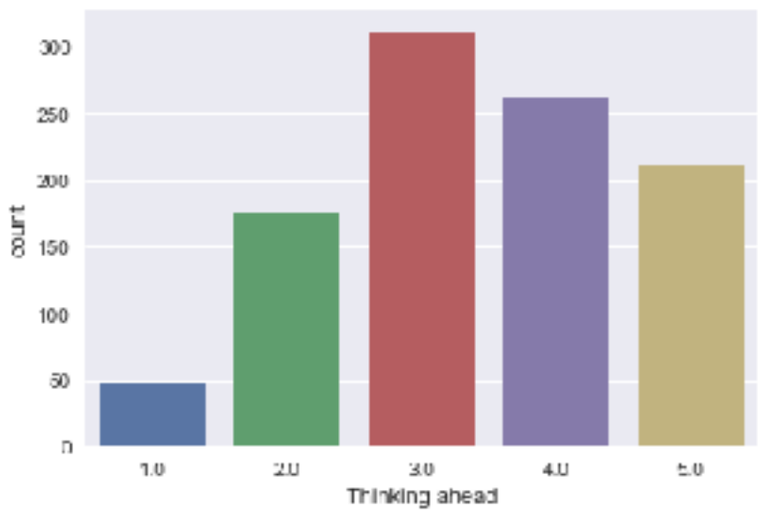
EXPLORING THE DATA

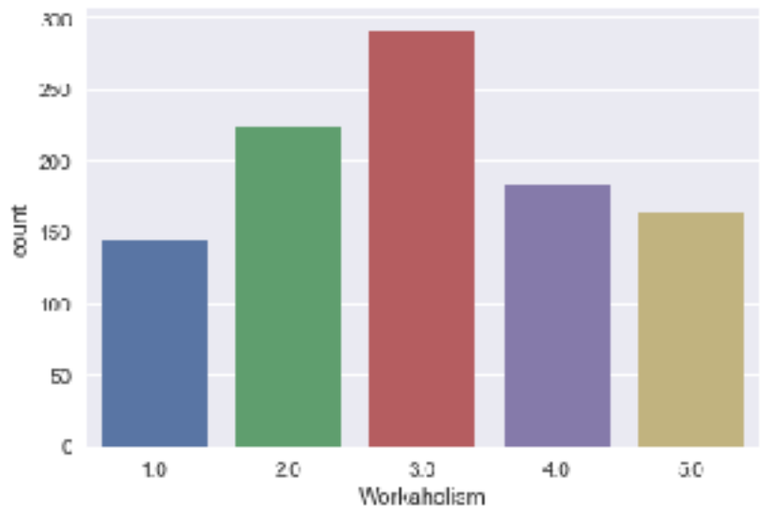
Skewness = -0.14

- Independent variables distribution:

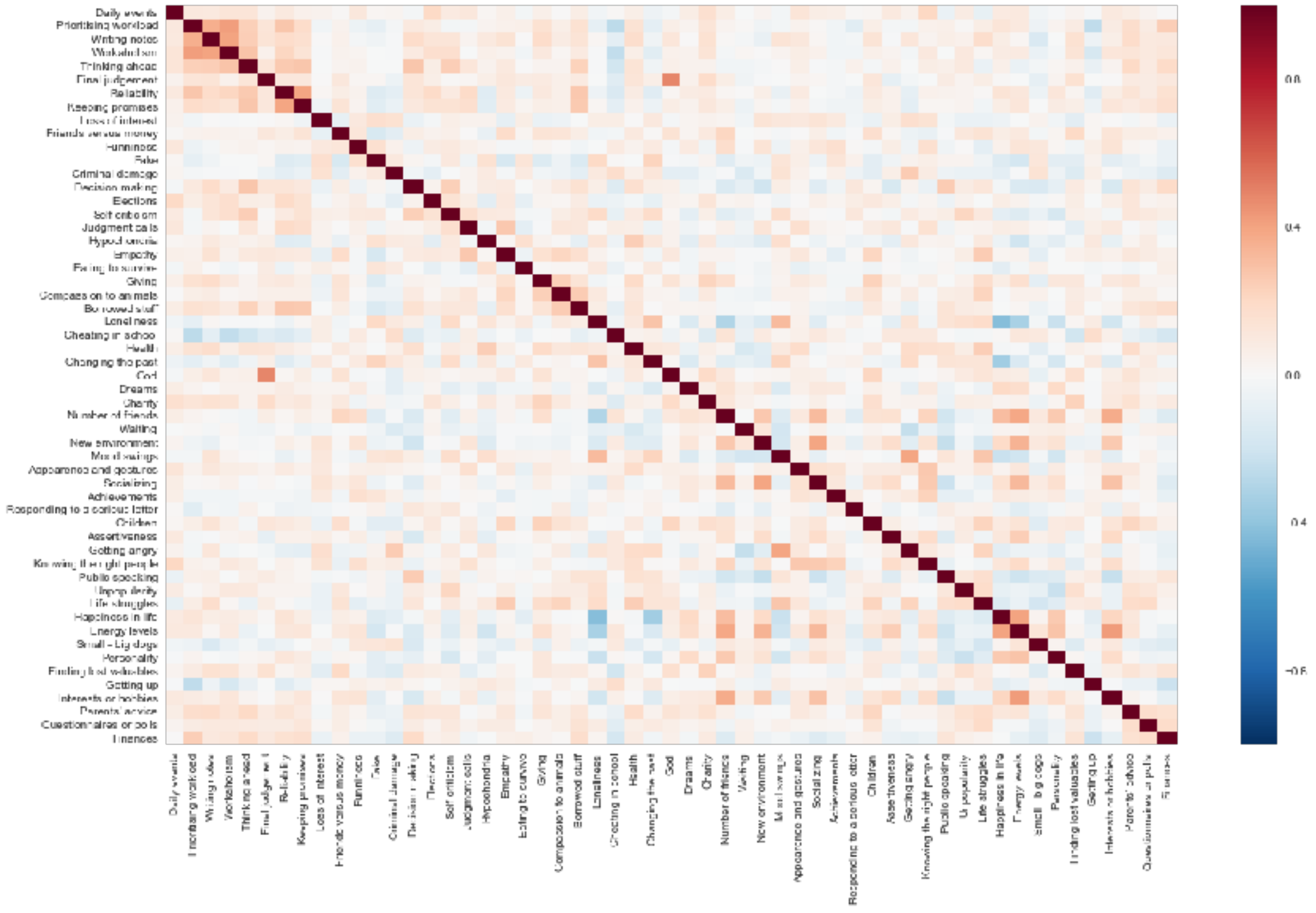








EXPLORING THE DATA



EXPLORING THE DATA

- Variables with positive correlation

God	Final judgement	0.488291
Final judgement	God	0.488291
Energy levels	Happiness in life	0.440593
Happiness in life	Energy levels	0.440593
Interests or hobbies	Energy levels	0.431741
Energy levels	Interests or hobbies	0.431741
Workaholism	Prioritising workload	0.418900
Prioritising workload	Workaholism	0.418900
Socializing	New environment	0.409969
New environment	Socializing	0.409969
Workaholism	Writing notes	0.409363
Writing notes	Workaholism	0.409363

- Variables with negative correlation

Loneliness	Number of friends	-0.316977
Number of friends	Loneliness	-0.316977
Energy levels	Loneliness	-0.346972
Loneliness	Energy levels	-0.346972
Changing the past	Happiness in life	-0.352469
Happiness in life	Changing the past	-0.352469
Loneliness	Happiness in life	-0.437061
Happiness in life	Loneliness	-0.437061

EXPLORING THE DATA

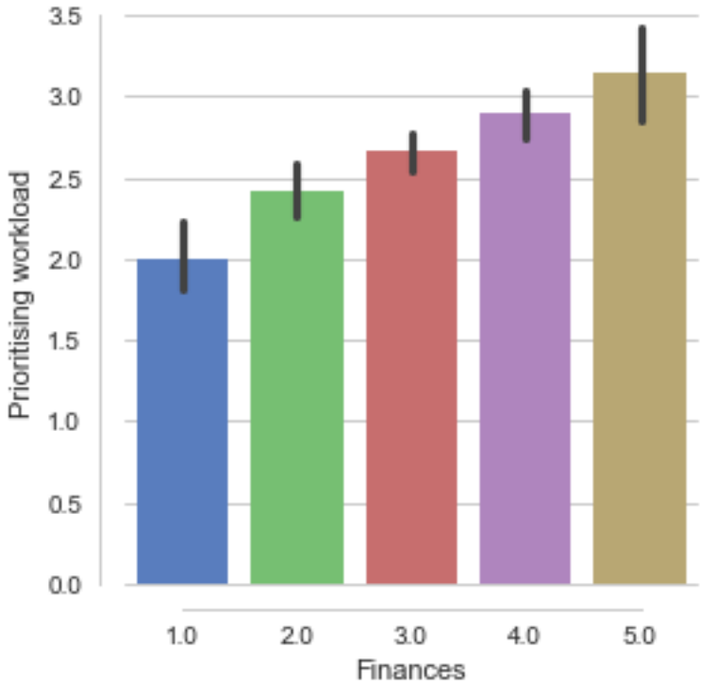
- Variables with positive correlation with target

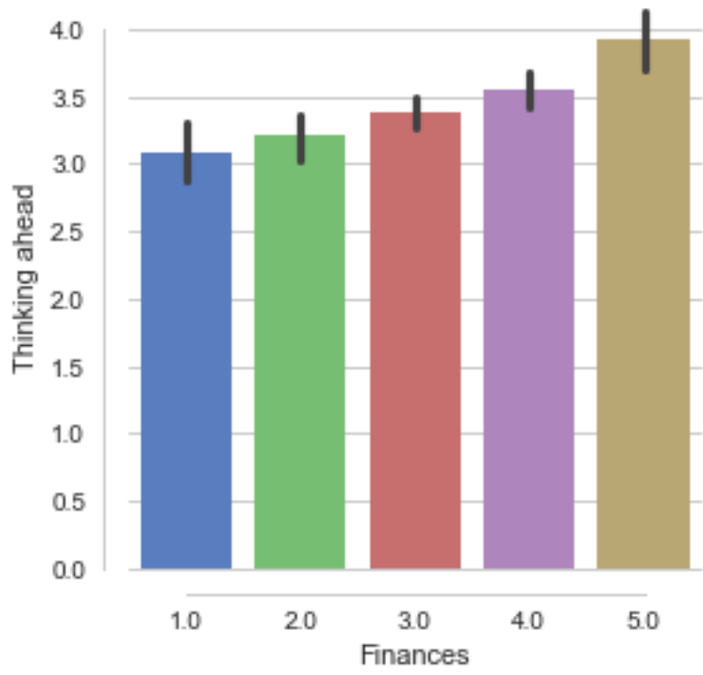
Prioritising workload	0.247062
Thinking ahead	0.190752
Decision making	0.190424
Questionnaires or polls	0.184353
Borrowed stuff	0.179615
Keeping promises	0.164246
Parents' advice	0.154827
Reliability	0.153559
Finding lost valuables	0.138382
Public speaking	0.131567

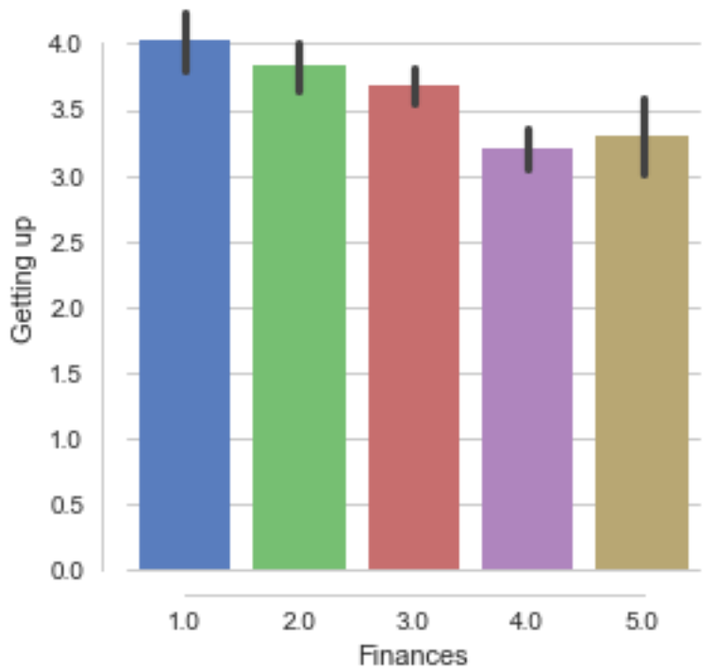
- Variables with negative correlation with target

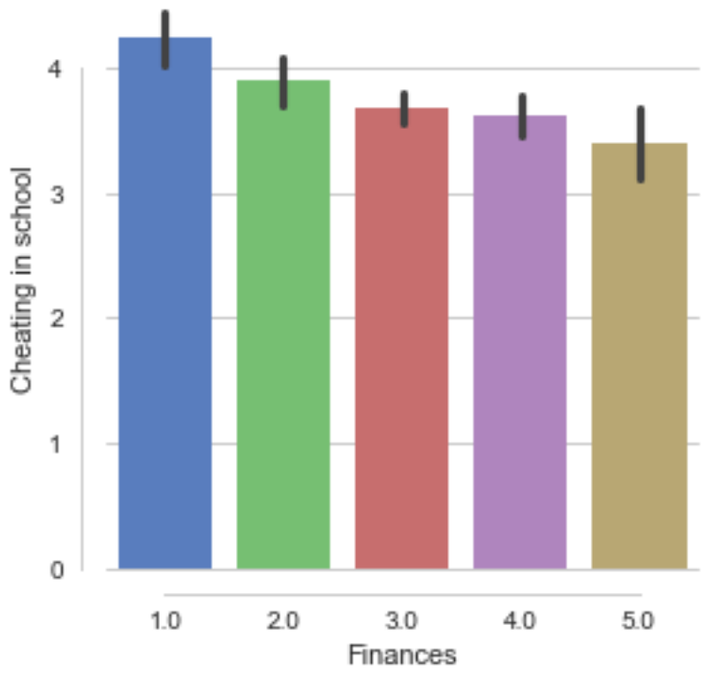
Achievements	-0.078346
Energy levels	-0.080409
Getting angry	-0.081181
Interests or hobbies	-0.085822
Socializing	-0.126491
Small - big dogs	-0.128435
Criminal damage	-0.128909
Number of friends	-0.137887
Cheating in school	-0.173914
Getting up	-0.213104

EXPLORING THE DATA









DATA PREPARATION

- Fill null values with column mode;
- Get dummies for necessary columns;
- Transformed label into binomial;
- Slip the data using train_test_split function.

```
df_data_model.loc[df_data_model['Finances'] <= 3, 'Finances'] = 0
df_data_model.loc[df_data_model['Finances'] > 3, 'Finances'] = 1
```

CHOICE OF METRIC

- Business objective: Bank will only give credit if the customer has a good spending habit (class 1).

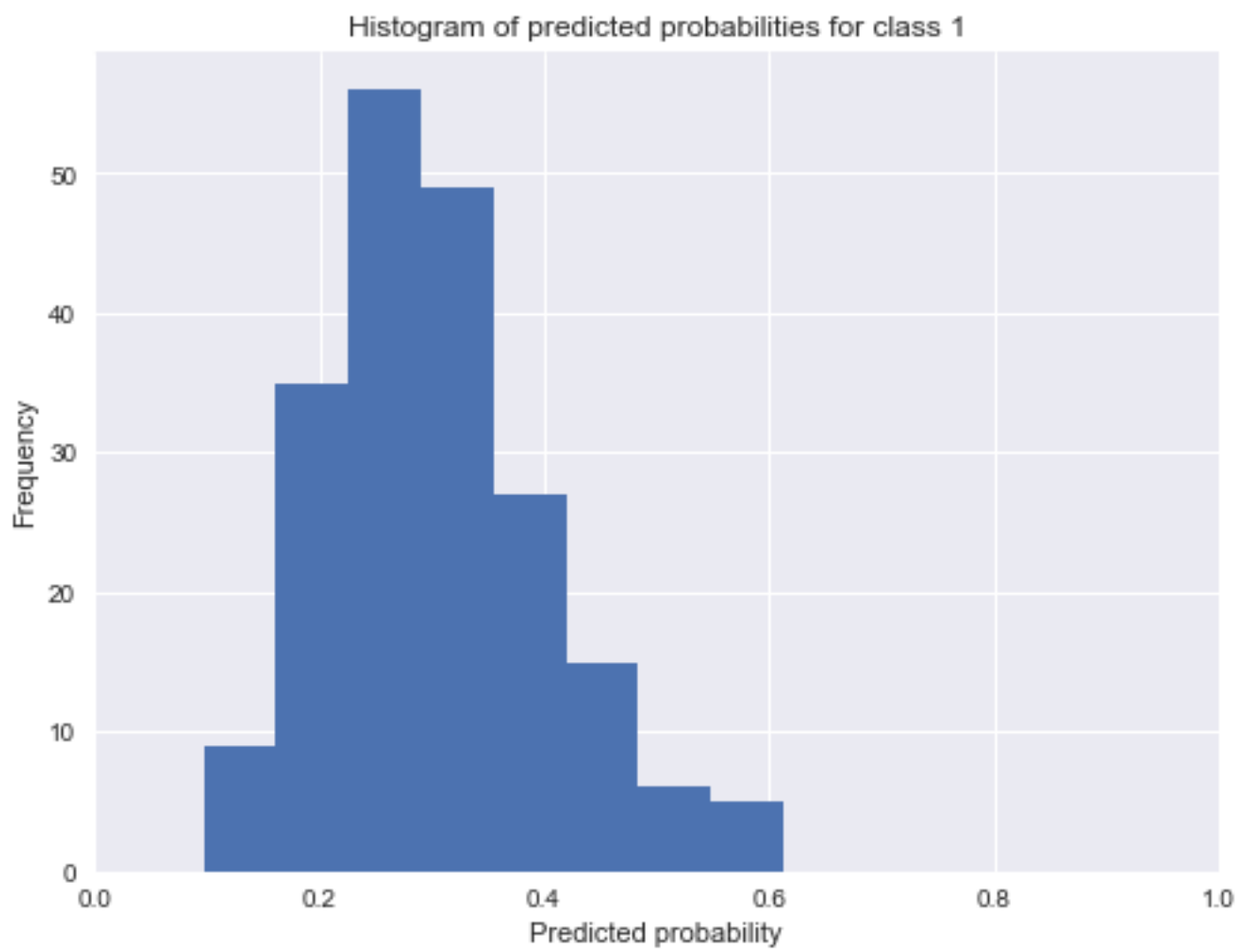
- For this scenario, the model should be optimized for **Precision** because it is **better** for the bank to not give credit to someone with a good spending habit (**False negative**) than to give credit to someone with a bad spending habit (**False positive**).

KNN MODEL – RESULTS

```
KNeighborsClassifier(algorithm='kd_tree', leaf_size=11, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=31, p=2,
           weights='uniform')
```

	Predicted: 0	Predicted :1
Actual: 0	129	4
Actual: 1	66	3

	precision	recall	f1-score	support
0.0	0.66	0.97	0.79	133
1.0	0.43	0.04	0.08	69
avg / total	0.58	0.65	0.54	202



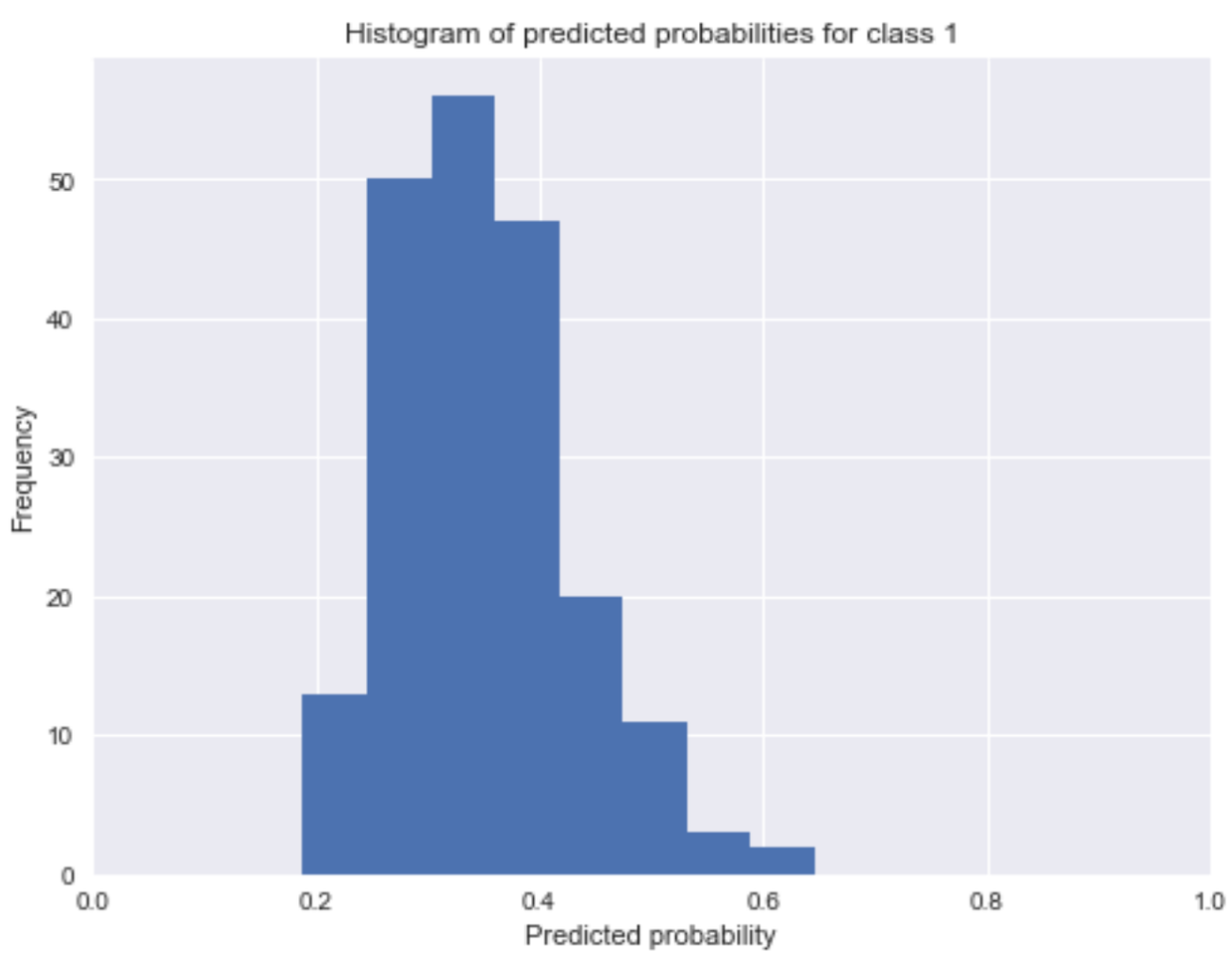
RANDOM FOREST MODEL – RESULTS

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=6, max_features=4, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            n_estimators=400, n_jobs=1, oob_score=False, random_state=None,
            verbose=0, warm_start=False)
```

	Features	Importance Score
50	Getting up	0.019332
1	Prioritising workload	0.014702
4	Thinking ahead	0.033274
63	Questionnaires or polls	0.030991
42	Public speaking	0.027762
7	Keeping promises	0.027088
3	Workaholism	0.022262
36	Achievements	0.021940
35	Socializing	0.021521
30	Number of friends	0.021455

	Predicted: 0	Predicted :1
Actual: 0	129	4
Actual: 1	63	6

	precision	recall	f1-score	support
0.0	0.67	0.97	0.79	133
1.0	0.60	0.09	0.15	69
avg / total	0.65	0.67	0.57	202

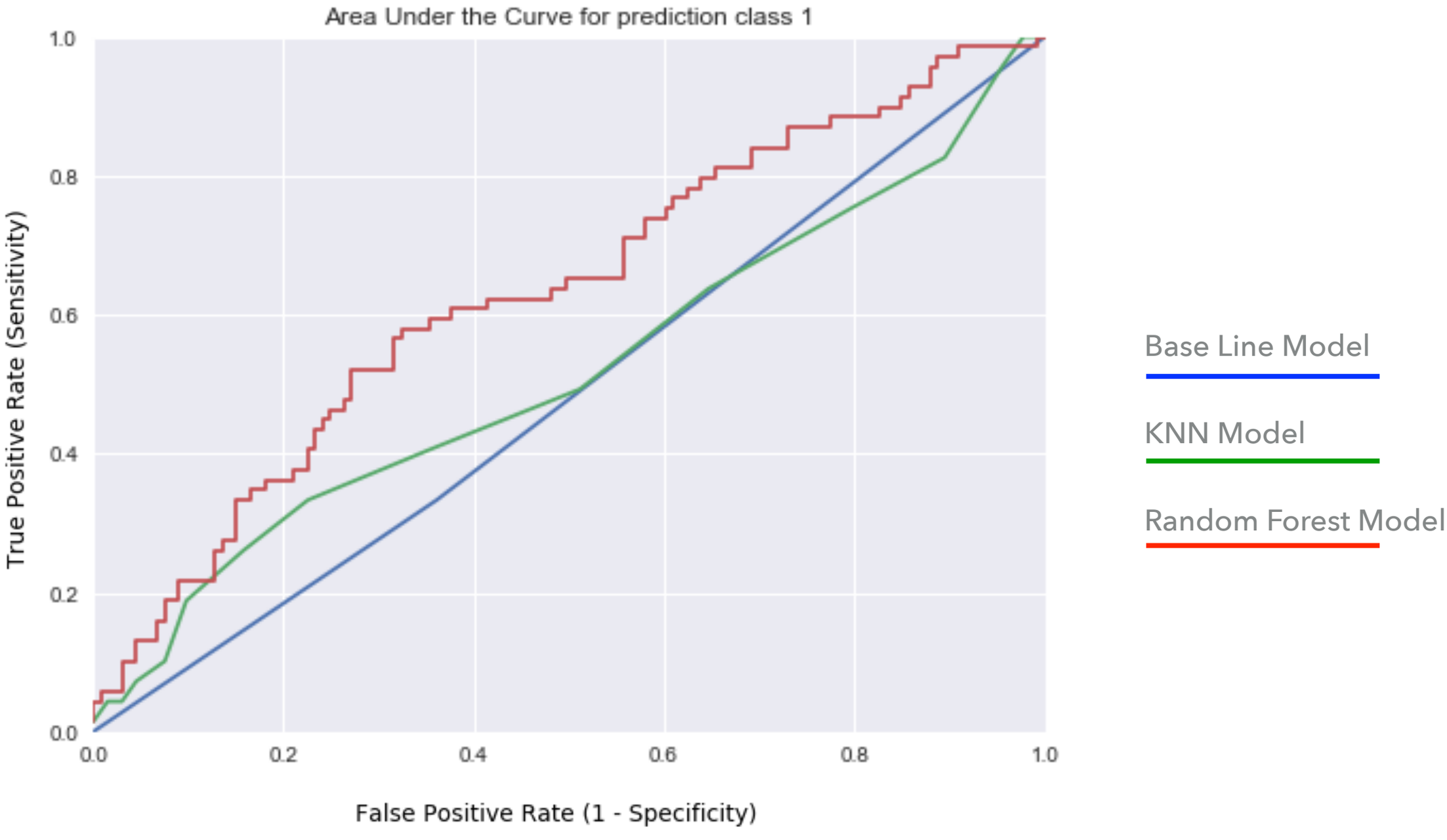


BASE LINE MODEL- RESULTS

	Predicted: 0	Predicted :1
Actual: 0	80	53
Actual: 1	47	22

	precision	recall	f1-score	support
0.0	0.63	0.60	0.62	133
1.0	0.29	0.32	0.31	69
avg / total	0.51	0.50	0.51	202

COMPARING MODELS



FUTURE WORK...

- Understand the influence of other variable groups (demographics, phobias, etc.) into spending habits.

- Apply the same survey to other age range participants and compare the results.

- Given the personality trait views on life & opinion, do people make up any clusters of similar behavior?

THANK YOU!