Задача

Предсказать категорию преступления по координатам, времени и другим признакам. (Соревнование на Kaggle - https://www.kaggle.com/c/sf-crime).

Описание данных

Данные включают информацию о преступлениях, совершенных в Сан Франциско в период с 1 января 2003 г. по 13 мая 2015 г. Данные предоставлены Департаментом полиции Сан-Франциско.
Данные включают:
Dates – дата и время совершения преступления
Category – тип преступления (указано только в обучающей выборке)
Descript – подробное описание преступления (указано только в обучающей выборке)
DayOfWeek – день недели совершения преступления
PdDistrict – название района, к которому принадлежит отделение полиции, расследовавшее данное преступление
Resolution – чем завершилось расследование (указано только в обучающей выборке)
Address – улица, где совершено преступление
X - долгота
Y – широта

Преступления и их количество:

LARCENY/THEFT	174900
OTHER OFFENSES	126182
NON-CRIMINAL	92304
ASSAULT	76876
DRUG/NARCOTIC	53971
VEHICLE THEFT	53781
VANDALISM	44725
WARRANTS	42214
BURGLARY	36755
SUSPICIOUS OCC	31414
MISSING PERSON	25989
ROBBERY	23000
FRAUD	16679
FORGERY/COUNTERFEITING	10609
SECONDARY CODES	9985
WEAPON LAWS	8555
PROSTITUTION	7484
TRESPASS	7326
STOLEN PROPERTY	4540
SEX OFFENSES FORCIBLE	4388
DISORDERLY CONDUCT	4320
DRUNKENNESS	4280
RECOVERED VEHICLE	3138
KIDNAPPING	2341
DRIVING UNDER THE INFLUENCE	2268
RUNAWAY	1946

LIQUOR LAWS	1903
ARSON	1513
LOITERING	1225
EMBEZZLEMENT	1166
SUICIDE	508
FAMILY OFFENSES	491
BAD CHECKS	406
BRIBERY	289
EXTORTION	256
SEX OFFENSES NON FORCIBLE	148
GAMBLING	146
PORNOGRAPHY/OBSCENE MAT	22
TREA	6

Обучающая выборка состоит из данных о преступлениях, совершенных в нечетные недели, тестовая выборка – в четные.
Размер данных:
Обучающая выборка – 127,4 МБ
Тестовая выборка – 91 МБ

Анализ и первоначальная обработка данных.

В обучающей выборке:
878 049 – объектов
8 признаков: Dates, Descript, DayOfWeek, PdDistrict, Resolution, Address, X, Y

При первой обработке данных были извлечены в отдельные признаки:

- Из Dates в Year, Month, Time (по часам).
- Удалено описание преступлений (Descript) и исход расследования (Resolution).

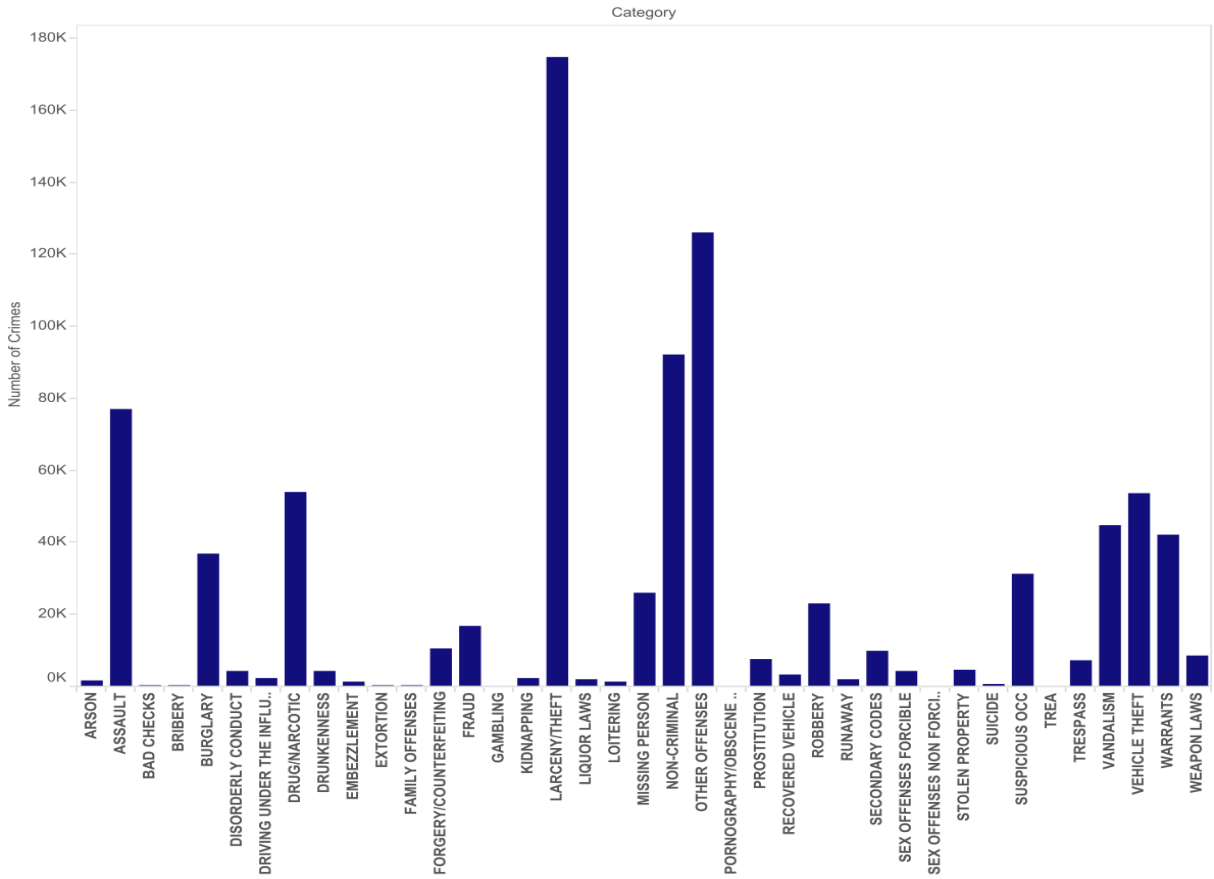
После первой обработки данных:
878 049 – объектов
9 признаков: Year, Month, Time, DayOfWeek, PdDistrict, Address, X, Y.

Статистика и визуализация:

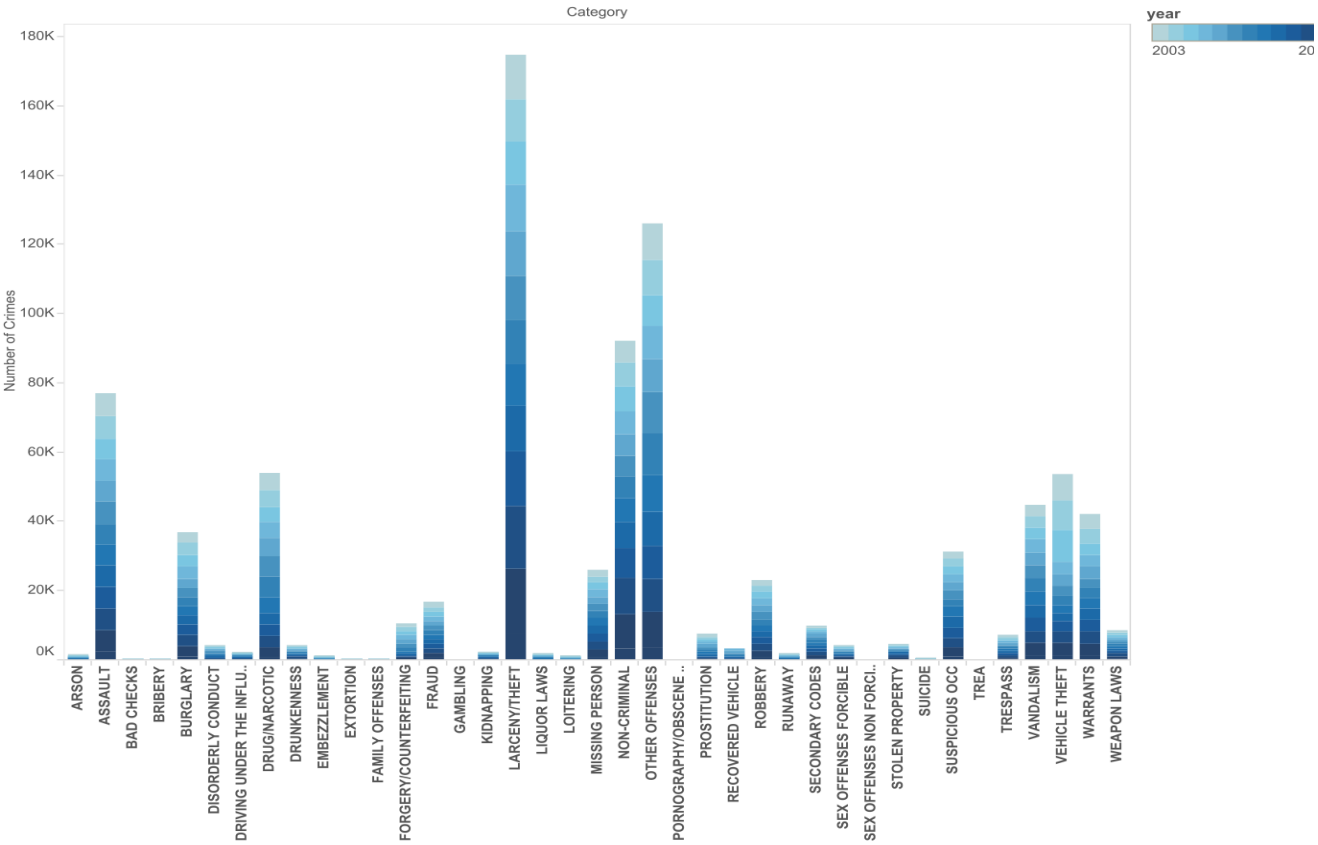
Карта преступлений, где можно выбрать категорию преступления, год, месяц, день недели, время суток, район и исход расследования.
https://public.tableau.com/views/SanFranciscoCrimes/Story1?:embed=y&:display_count=yes&:showTabs=y

Карта преступлений сделана в Tableau, графики – в Tableau и R.
Ниже приведены 6 графиков, наиболее ярко отражающих криминальную ситуацию в Сан Франциско, и выводы, основанные на этих графиках и карте преступлений.

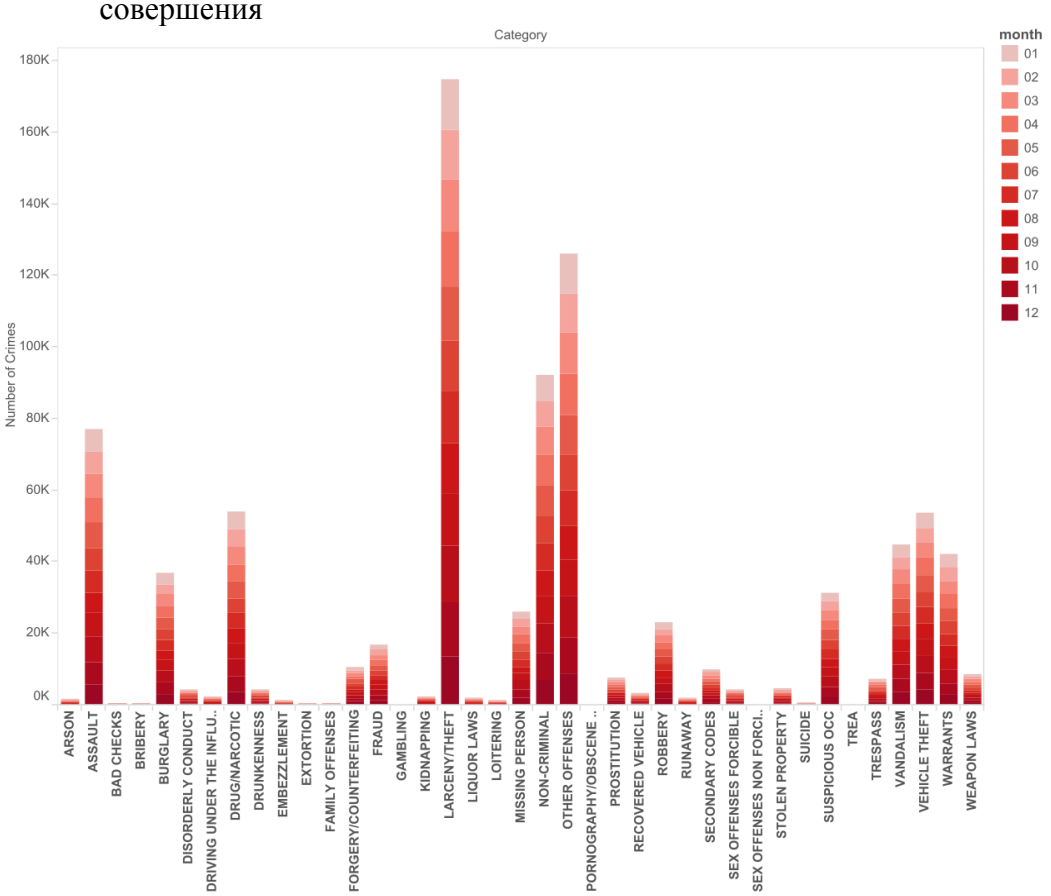
1. Количество преступлений на каждую категорию.



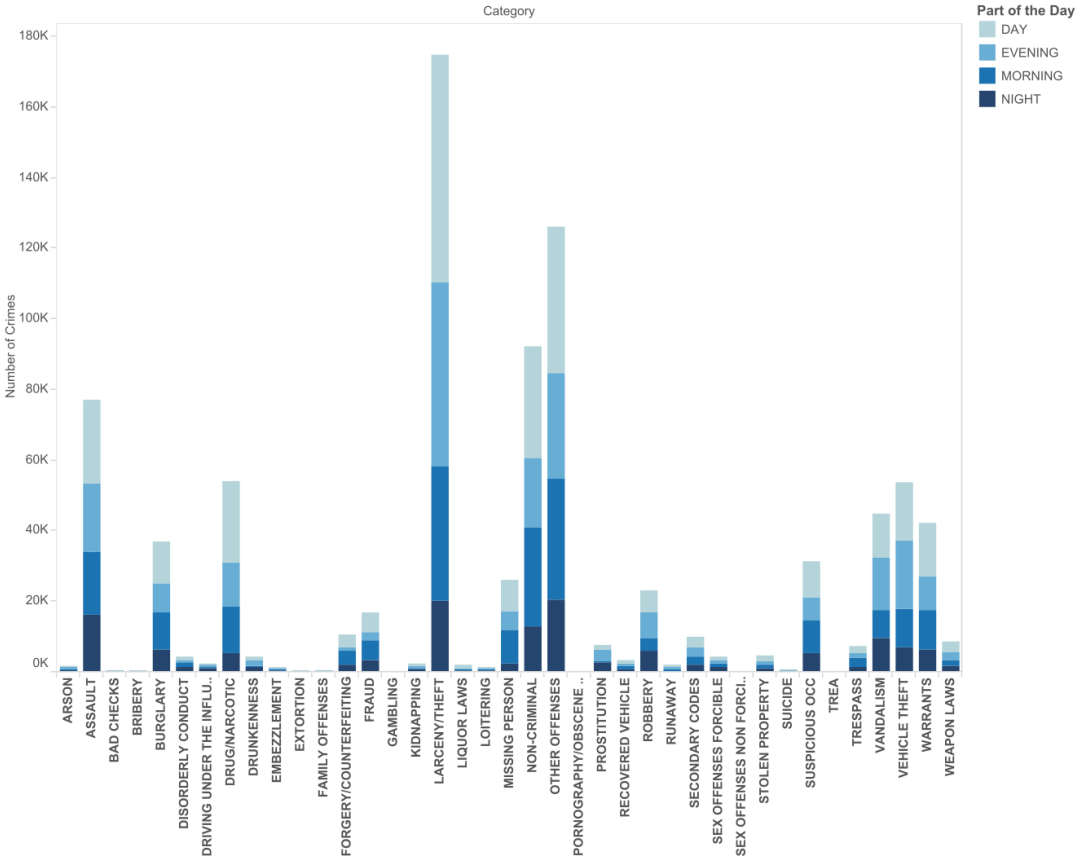
2. Количество преступлений на каждую его категорию в зависимости от года его совершения.



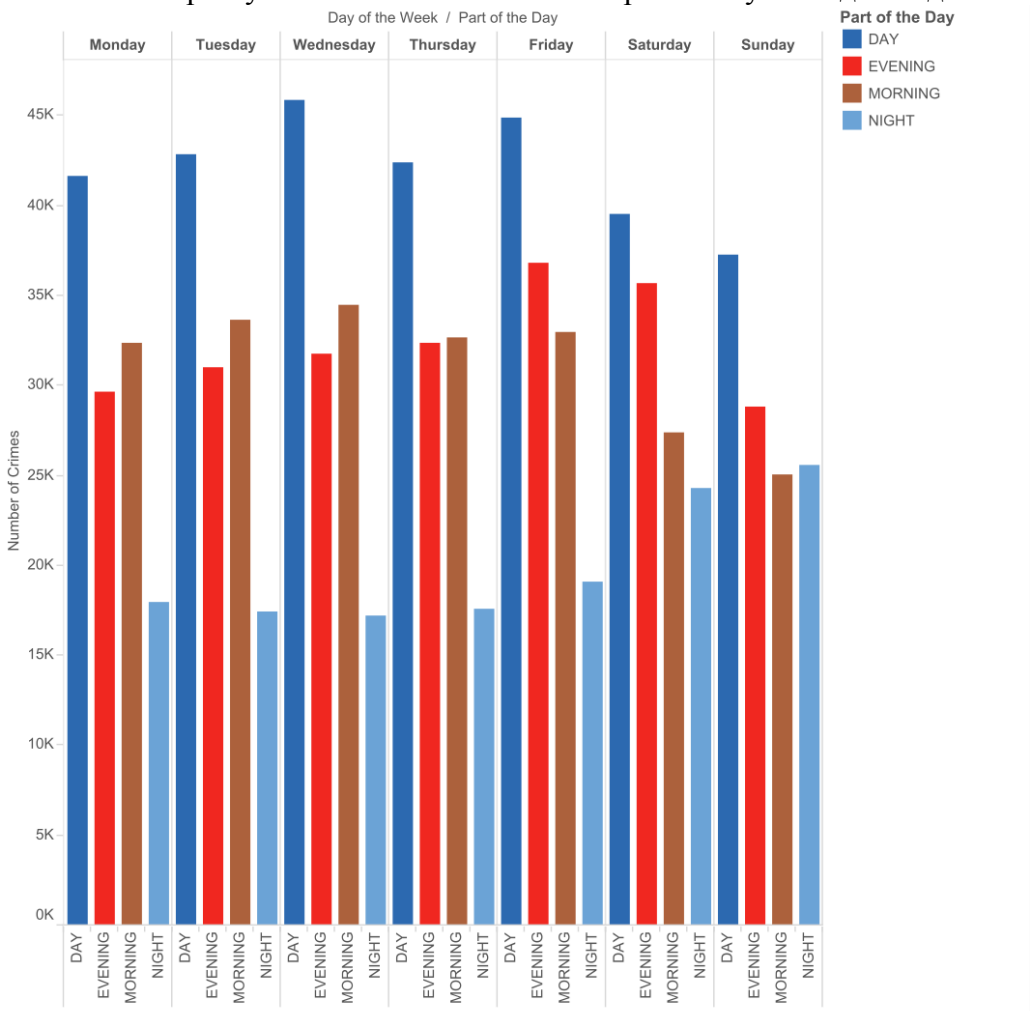
3. Количество преступлений на каждую его категорию в зависимости от месяца его совершения



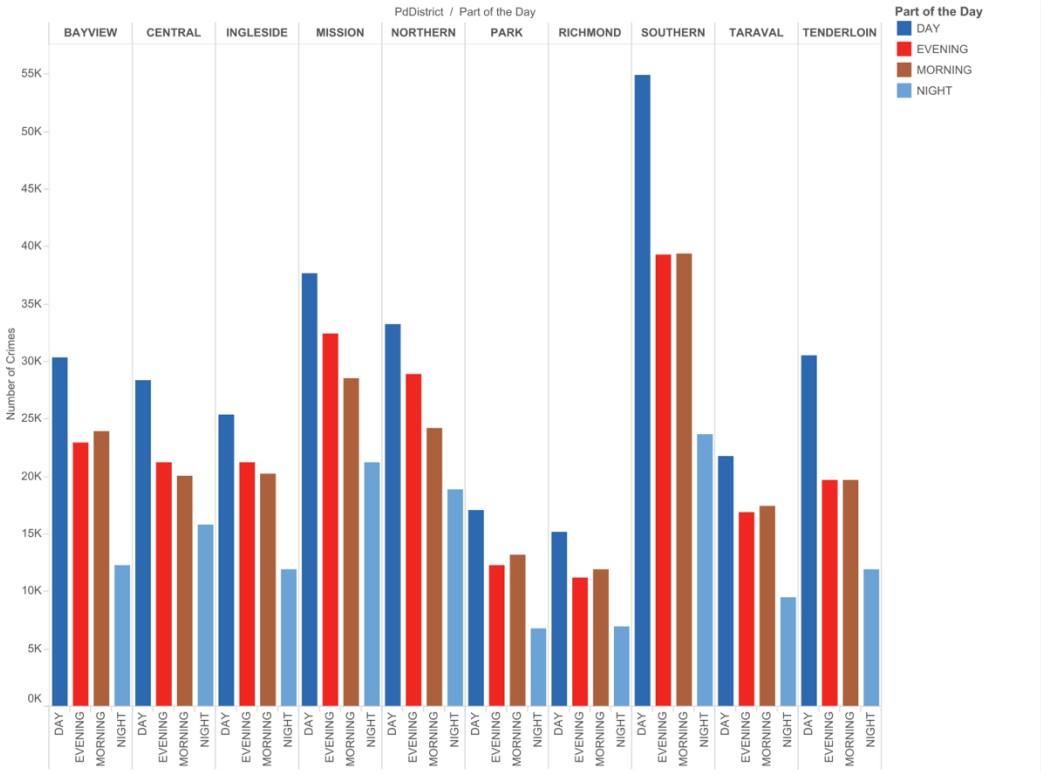
4. Количество преступлений на каждую его категорию в зависимости от времени суток его совершения



5. Количество преступлений в зависимости от времени суток и дня недели.



6. Количество преступлений в зависимости от времени суток и района, в котором совершилось преступление.



Выводы по карте и графикам:

1. Топ 5 преступлений: LARCENY/THEFT, OTHER OFFENSES и NON-CRIMINAL, ASSAULT, DRUG/NARCOTIC.
2. Преступления в зависимости от года их совершения распределены равномерно.
3. Та же ситуация и с распределением преступлений по месяцам – они распределены равномерно.
4. Преступления по дням недели тоже распределены достаточно равномерно.
5. В целом распределение преступлений в зависимости от времени суток равномерно. Днем совершается большее количество преступлений, ночью - меньшее. Но тенденция меняется относительно дней недели. Так, например, в пятницу и субботу преступлений, совершенных вечером, становится больше.
6. Преступления плотно распределены по всему городу. Это ярко отражает карта преступлений, визуализированная в Tableau.
7. Тенденция распределения преступлений по районам сохраняется из года в год и из месяца в месяц. При этом по количеству совершенных преступлений лидирует район Southern (независимо от времени суток). На втором и третьем месте районы Mission и Northern соответственно.

Ход выполнения.

Первый эксперимент.

Изначально мы хотели загрузить наши данные в Weka и Orange, обработать их, запустить несколько различных классификаторов и посмотреть на результаты.

Обработка в Weka:
Удалили признак Address.
Применили NumericToNominal для Year, Month, Time
Применили Center для признаков X,Y.
Применили NominalToBinary для признаков Year, Month, Time, DayOfWeek, PdDistrict.

После обработки:
878 049 – объектов
48 признаков

Данные оказались слишком большими для Weka, и нам не удалось запустить ни один классификатор на всех данных. Поэтому с помощью фильтра Resample мы взяли 30% от наших данных. Получилось следующее:

Данные:
263 414 объектов
48 признаков

Результаты в Weka:

1. **OneR**.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Weighted Avg.	0,261	0,122	0,215	0,261	0,209	0,131	0,569	0,131

2. **NaïveBayes** на кросс-валидации 3.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Weighted Avg.	0,144	0,081	0,154	0,144	0,106	0,052	0,600	0,133

Остальные классификаторы не удалось запустить в Weka и мы решили попробовать Orange с этой же выборкой. Orange работал немного получше и не ругался на недостаток памяти. Однако с результатами все равно все плохо.

Результаты в Orange:

1. **Classification Tree** на кросс-валидации 5.

Method	AUC	CA	F1	Precision	Recall
Classification Tree	0.567	0.264	0.207	0.209	0.264

2. **Logistic Regression** на кросс-валидации 5.

Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.534	0.228	0.139	0.173	0.228

По итогам первого эксперимента стало понятно, что просто не будет.
Данные оказались слишком большими для обеих программ, поэтому мы удалили из данных самые многочисленные преступления – LARCENY/THEFT, OTHER OFFENSES и NON-CRIMINAL.

Результаты классификаторов оказались ужасно плохими (стыдно показывать), поэтому мы решили отказаться от классов, которые предсказываются хуже всего – TREA, PORNOGRAPHY/OBSCENE MAT, GAMBLING, BRIBERY и EXTORTION.

Результаты в Weka:

1. **OneR** на кросс-валидации 3.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Weighted Avg.	0,248	0,100	0,218	0,248	0,207	0,139	0,574	0,127

2. **NaïveBayes** на кросс-валидации 3.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Weighted Avg.	0,194	0,063	0,242	0,194	0,142	0,123	0,678	0,194

Результаты в Orange:

1. **Classification Tree** на кросс-валидации 5.

Method	AUC	CA	F1	Precision	Recall
Classification Tree	0.622	0.332	0.293	0.297	0.332

2. **Logistic Regression** на кросс-валидации 5.

Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.619	0.328	0.290	0.315	0.328

Второй эксперимент.

Мы отказались от Orange и Weka и стали работать в sklearn.
Прежде чем писать какой-то код нам надо было установить, чем вызваны такие плохие результаты. Поэтому мы стали подробнее изучать данные (сделали визуализацию в tableau), почитали блог kaggle и форум самого соревнования.

Из визуализации стало понятно, что точных предсказаний сделать практически невозможно, так как преступления (за редким исключением) плотно распределены на территории города, поэтому и оценка в первом эксперименте такая низкая.
В sklearn есть такая возможность и подходящий метод оценивания – logloss. В нескольких решениях, выложенных на форуме мы как раз обнаружили эти методы и узнали, что kaggle оценивает результаты именно так.
Мы написали небольшой скрипт, который проводил такие же действия с данными как и в первом эксперименте и обучили на них логистическую регрессию.

Данные:
878 049 – объектов
48 - признака

Результаты:
Обучающая выборка – 2.55
Тестовая выборка – 2.55

Результаты оказались довольно хорошие. Такой оценке примерно соответствует 500 место на kaggle.

Третий эксперимент.

Несколько других простых классификаторов дали нам не такие хорошие результаты, поэтому мы стали думать, как можно задействовать признак ‘Address’, который мы до этого просто выкидывали.
Адрес в данных указывался с точностью до номера дома, что делает большую часть наблюдений уникальными, поэтому мы удалили все кроме названий улиц. Получилось около 14 тыс. уникальных адресов.
Большая их часть, однако, приходится на перекрестки. Количество уникальных улиц в итоге оказалось равно 2 тысячам. Это все равно очень много. Мы попробовали применить CountVectorizer, но получившийся файл весил 3.7 гб. Пробовать делать, что-то с такими объемами мы не решились.
Решение мы нашли в одном из скриптов на kaggle. Суть решения в том, чтобы использовать статистику, собранную на всех данных. В частности для каждого адреса вычисляется логарифм отношения шансов(logodds), для пересечения вводится один дополнительный бинарный признак. В итоге получается 39 дополнительных признаков. В итоге получили следующее:

Данные:
878 049 – объектов
63 признака

Результаты:
Обучающая выборка – 2.21401473385
Тестовая выборка – 2.21948940207

Автор этого скрипта, однако, не бинаризировал остальные признаки (месяц, дату, год), а время у него кодировалось лишь одним признаком ‘Awake’. Мы расширили количество признаков времени до 4(день, ночь, вечер, утро), а также добавили признаки из нашего второго эксперимента. Результат:
Данные:
878 049 – объектов
121 признак

Результаты:
Обучающая выборка - 2.20397602055
Тестовая выборка 2.2097351364

Придумать еще каких-то несложных улучшений у нас не получилось, и мы решили остановиться на этом результате. Если предположить, что kaggle оценит нас также, то мы попадем в первую десятку.