**Подготовка данных и обработка**

В качестве исходных данных были взяты статистические данные компании Lending Club по выданным ссудам за 2014 год. Отчеты за 2014 год были выбраны по причине того, что у большинства ссуд, выданных позднее еще не наступил срок погашения. Статистическая совокупность имеет 235 629 записей о выданных ссуда. Каждая запись является носителем 135 признаков. Из которых 134 свободных и один целевой - оценка заемщика. Среди признаков встречаются как числовые, так и атрибутивные признаки.

Первичный осмотр данных показал, что данные содержат существенное количество пропусков, пустых признаков, а также то, что атрибутивные признаки представлены в строчном формате. Перед тем как передать данные нейронной сети, необходимо провести подготовку данных.

Во-первых, удалим все признаки, которые не содержат ни одного значения. От такого действия никакая существенная информация априори не может быть потеряна. В результате были удалены следующие 29 признаков: url , annual\_inc\_joint , dti\_joint , verification\_status\_joint , open\_acc\_6m , open\_il\_6m , open\_il\_12m , open\_il\_24m , mths\_since\_rcnt\_il , total\_bal\_il , il\_util , open\_rv\_12m , open\_rv\_24m , max\_bal\_bc , all\_util , inq\_fi , total\_cu\_tl , inq\_last\_12m , revol\_bal\_joint , sec\_app\_earliest\_cr\_line , sec\_app\_inq\_last\_6mths , sec\_app\_mort\_acc , sec\_app\_open\_acc , sec\_app\_revol\_util , sec\_app\_open\_il\_6m , sec\_app\_num\_rev\_accts , sec\_app\_chargeoff\_within\_12\_mths , sec\_app\_collections\_12\_mths\_ex\_med , sec\_app\_mths\_since\_last\_major\_derog.

Опишем все переменные и их смысл.

|  |  |  |
| --- | --- | --- |
|  | loan\_amnt | Указанная сумма займа, поданного заемщиком. Если в какой-то момент кредитный отдел уменьшит сумму кредита, то это будет отражено в данном признаке. |
|  | funded\_amnt | Общая сумма, начисленная на этот кредит на данный момент времени. |
|  | funded\_amnt\_inv | Общая сумма, полученная инвесторами за этот кредит на тот момент. |
|  | term | Количество платежей по кредиту. Значения указаны в месяцах и могут принимать значения 36 или 60. |
|  | int\_rate | Процентная ставка по кредиту |
|  | installment | Ежемесячный платеж, который должен платить заемщик |
|  | grade | Оценка от LC |
|  | sub\_grade | Подробная оценка от LC |
|  | emp\_title | Должность, предоставленная Заемщиком при подаче заявки на получение кредита. |
|  | emp\_length | Занятость в годах. Возможные значения: от 0 до 10, где 0 означает менее одного года, а 10 означает десять или более лет. |
|  | home\_ownership | Статус собственности на дом, предоставляемый заемщиком во время регистрации или полученный из кредитного отчета. Возможные значения: АРЕНДА, СОБСТВЕННОСТЬ, ИПОТЕКА, ДРУГОЕ. |
|  | annual\_inc | Годовой доход, предоставляемый в анкете заемщиком во время регистрации. |
|  | verification\_status | Указывает, проверен ли уровень дохода заемщика |
|  | issue\_d | Месяц, в который был получен заём. |
|  | loan\_status | Текущий статус кредита |
|  | pymnt\_plan | Указывает, был ли установлен план платежей по кредиту |
|  | desc | Описание займа, предоставленное заемщиком |
|  | purpose | Категория, предоставленная заемщиком для запроса займа. |
|  | title | Название цели займа, предоставленное заемщиком |
|  | zip\_code | Первые 3 цифры почтового индекса, предоставленные заемщиком в заявке на получение кредита. |
|  | addr\_state | Штат, указанный заемщиком в заявке на получение кредита |
|  | dti | Соотношение, рассчитанное с использованием сумм ежемесячных платежей заемщика по общим долговым обязательствам, за исключением ипотечного кредита и запрошенного кредитного кредита, и разделенное на ежемесячный доход заемщика. |
|  | delinq\_2yrs | Число просроченных 30-дневных случаев делинквентности в кредитном деле заемщика за последние 2 года |
|  | earliest\_cr\_line | Месяц, когда была открыта самая ранняя кредитная линия заемщика |
|  | inq\_last\_6mths | Количество запросов за последние 6 месяцев (за исключением запросов на авто и ипотеку) |
|  | mths\_since\_last\_delinq | Число месяцев с момента последней просрочки платежа по задолженности заемщика. |
|  | mths\_since\_last\_record | Число месяцев с момента последнего публичной записи. |
|  | open\_acc | Количество открытых кредитных линий в кредитном деле заемщика. |
|  | pub\_rec | Количество нарушающих публичных записей |
|  | revol\_bal | Общий кредитный оборотный баланс |
|  | revol\_util | Ставка использования возобновляемых кредитных линий или сумма кредита, которую заемщик использует по отношению ко всему доступному возобновляемому кредиту. |
|  | total\_acc | Общее количество кредитных линий, находящихся в настоящее время в кредитном деле заемщика |
|  | initial\_list\_status | Первоначальный статус листинга по кредиту. Возможные значения: W, F |
|  | out\_prncp | Остаток непогашенной основной суммы на общую сумму |
|  | out\_prncp\_inv | Остаток непогашенной основной суммы для части общей суммы, финансируемой инвесторами |
|  | total\_pymnt | Выплаты, полученные на сегодняшний день для общей суммы, финансируемой инвесторами |
|  | total\_pymnt\_inv | Платежи, полученные до настоящего времени для части общей суммы, финансируемой инвесторами |
|  | total\_rec\_prncp | Доверитель, полученный до настоящего времени |
|  | total\_rec\_int | Проценты, полученные на сегодняшний день |
|  | total\_rec\_late\_fee | Полученные на сегодняшний день взносы |
|  | recoveries | после оплаты |
|  | collection\_recovery\_fee | Сделать проводку за сбор |
|  | last\_pymnt\_d | Месяц, когда был получен платеж |
|  | last\_pymnt\_amnt | Полученная сумма общего платежа |
|  | next\_pymnt\_d | Дата следующего планового платежа |
|  | last\_credit\_pull\_d | Последний месяц, когда LC привлек кредит для этого займа |
|  | collections\_12\_mths\_ex\_med | Количество сборов денег за 12 месяцев без учета медицинских сборов |
|  | mths\_since\_last\_major\_derog | Месяцы с момента последнего 90-дневного или худшего рейтинга |
|  | policy\_code | общедоступный (policy\_code = 1), новые продукты недоступные для общественности( policy\_code = 2) |
|  | application\_type | Указывает, является ли кредит индивидуальным или совместным |
|  | acc\_now\_delinq | Количество учетных записей, на которых заемщик является нарушителем условий ссуды |
|  | tot\_coll\_amt | Общая сумма сбора, которую задолжал заемщик |
|  | tot\_cur\_bal | Общий текущий баланс всех счетов |
|  | total\_rev\_hi\_lim | - |
|  | acc\_open\_past\_24mths | Количество сделок, открытых за последние 24 месяца |
|  | avg\_cur\_bal | Средний текущий баланс всех счетов |
|  | bc\_open\_to\_buy | Всего можно купить на сумму на кредитных картах, находящихся в обращении |
|  | bc\_util | Соотношение общего текущего баланса с высоким кредитным лимитом для всех банковских счетов. |
|  | chargeoff\_within\_12\_mths | Количество списаний в течение 12 месяцев |
|  | delinq\_amnt | Просроченная сумма, причитающаяся за счета, на которых заемщик теперь является правонарушителем. |
|  | mo\_sin\_old\_il\_acct | Месяцы с момента открытия первого банковского счета |
|  | mo\_sin\_old\_rev\_tl\_op | Месяцы с момента открытия самой старой действующей учетной записи |
|  | mo\_sin\_rcnt\_rev\_tl\_op | Месяцы с момента открытия последней действующей учетной записи |
|  | mo\_sin\_rcnt\_tl | Месяцы с момента открытия последней учетной записи |
|  | mort\_acc | Количество ипотечных счетов |
|  | mths\_since\_recent\_bc | Месяцы с момента открытия последней банковской карты. |
|  | mths\_since\_recent\_bc\_dlq | Месяцы с момента последней просрочки банковской карточки |
|  | mths\_since\_recent\_inq | Месяцы с момента последнего расследования. |
|  | mths\_since\_recent\_revol\_delinq | Месяцы с момента последней действующей просрочки. |
|  | num\_accts\_ever\_120\_pd | Количество счетов в течение 120 или более дней |
|  | num\_actv\_bc\_tl | Количество активных банковских карточек |
|  | num\_actv\_rev\_tl | Количество действующих в настоящее время сделок |
|  | num\_bc\_sats | Количество удовлетворительных банковских счетов |
|  | num\_bc\_tl | Количество счетов для банковских карт |
|  | num\_il\_tl | Количество учетных записей |
|  | num\_op\_rev\_tl | Количество открытых возобновляемых счетов |
|  | num\_rev\_accts | Количество возобновляемых счетов |
|  | num\_rev\_tl\_bal\_gt\_0 | Количество возобновляемых сделок с положительным балансом |
|  | num\_sats | Количество удовлетворительных счетов |
|  | num\_tl\_120dpd\_2m | Количество текущих счетов за 120 дней (обновлено за последние 2 месяца) |
|  | num\_tl\_30dpd | Количество учетных записей, которые в настоящее время просрочены на 30 дней (обновлено за последние 2 месяца) |
|  | num\_tl\_90g\_dpd\_24m | Количество счетов в течение 90 и более дней, просроченных в последние 24 месяца |
|  | num\_tl\_op\_past\_12m | Количество счетов, открытых за последние 12 месяцев |
|  | pct\_tl\_nvr\_dlq | Процент сделок никогда не был просроченным |
|  | percent\_bc\_gt\_75 | Процент всех банковских счетов> 75% от лимита. |
|  | pub\_rec\_bankruptcies | Количество публичных банкротств |
|  | tax\_liens | Количество налоговых залогов |
|  | tot\_hi\_cred\_lim | Общий кредитный лимит |
|  | total\_bal\_ex\_mort | Общий кредитный баланс без учета ипотеки |
|  | total\_bc\_limit | Общий кредитный лимит банковской карты |
|  | total\_il\_high\_credit\_limit | Максимальный кредитный / кредитный лимит |
|  | hardship\_flag | Фискальный знак |
|  | hardship \_type | Тип тяжести |
|  | hardship\_reason | Причина тяжести |
|  | hardship\_status | Состояние тяжести |
|  | deferral\_term | Отсрочка |
|  | hardship\_amount | Количество трудностей |
|  | hardship\_start\_date | Дата начала тяжелой работы |
|  | hardship\_end\_date | Крайняя дата окончания |
|  | payment\_plan\_start\_date | Дата начала плана платежей |
|  | hardship\_length | Продолжительность тяги |
|  | hardship\_dpd | Количество дней просрочки по кредиту |
|  | hardship\_loan\_status | Состояния займа ( трудности) |
|  | orig\_projected\_additional\_accrued\_interest | Дополнительные начисленные проценты |
|  | hardship\_payoff\_balance\_amount | Сумма остатка средств |
|  | hardship\_last\_payment\_amount | Сумма невыплат |

Удалим признаки, которыми не может располагать инвестор на момент рассмотрения заявки. К числу таких признаков относятся: funded\_amnt\_inv, out\_prncp, out\_prncp\_inv, total\_pymnt, total\_pymnt\_inv, total\_rec\_prncp, total\_rec\_int, total\_rec\_late\_fee, recoveries, collection\_recovery\_fee, last\_pymnt\_d, last\_pymnt\_amnt, next\_pymnt\_d, last\_credit\_pull\_d, total\_rev\_hi\_lim, total\_il\_high\_credit\_limit, hardship\_flag, hardship \_type, hardship\_reason, hardship\_status, deferral\_term, hardship\_amount, hardship\_start\_date, hardship\_end\_date, hardship\_length, hardship\_dpd, hardship\_loan\_status, orig\_projected\_additional\_accrued\_interest, hardship\_payoff\_balance\_amount, hardship\_last\_payment\_amount, payment\_plan\_start\_date.

Создадим описание оставшихся признаков. В описании для каждого признака будет указан тип переменной, количество пропусков, доля пропусков, количество уникальных элементов, доля уникальных элементов, среднее значение, медиана, стандартное отклонение, максимум и минимум. Также для каждого числового признака построим гистограммы распределения.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| переменные | тип переменной | доля пропусков, % | доля уникальных элементов, % | среднее значение | медиана | стандартное отклонение | максимум | минимаум |
| loan\_amnt | целочисленный | 0.00 | 0.56 | 14870.157 | 13000.000 | 8438.318 | 35000.000 | 1000.000 |
| funded\_amnt | целочисленный | 0.00 | 0.56 | 14870.157 | 13000.000 | 8438.318 | 35000.000 | 1000.000 |
| term | атрибутивный | 0.00 | 0.00 | - | - | - | - | - |
| int\_rate | вещественный | 0.00 | 0.04 | 0.138 | 0.137 | 0.043 | 0.261 | 0.060 |
| installment | вещественный | 0.00 | 14.18 | 442.463 | 384.060 | 245.044 | 1409.990 | 23.360 |
| grade | атрибутивный | 0.00 | 0.00 | - | - | - | - | - |
| sub\_grade | атрибутивный | 0.00 | 0.01 | - | - | - | - | - |
| emp\_title | атрибутивный | 5.62 | 31.98 | - | - | - | - | - |
| emp\_length | атрибутивный | 0.00 | 0.01 | - | - | - | - | - |
| home\_ownership | атрибутивный | 0.00 | 0.00 | - | - | - | - | - |
| annual\_inc | вещественный | 0.00 | 7.74 | 74854.148 | 65000.000 | 55547.533 | 7500000.000 | 3000.000 |
| verification\_status | атрибутивный | 0.00 | 0.00 | - | - | - | - | - |
| issue\_d | атрибутивный | 0.00 | 0.01 | - | - | - | - | - |
| loan\_status | атрибутивный | 0.00 | 0.00 | - | - | - | - | - |
| pymnt\_plan | атрибутивный | 0.00 | 0.00 | - | - | - | - | - |
| desc | атрибутивный | 93.52 | 6.30 | - | - | - | - | - |
| purpose | атрибутивный | 0.00 | 0.01 | - | - | - | - | - |
| title | атрибутивный | 0.00 | 0.87 | - | - | - | - | - |
| zip\_code | атрибутивный | 0.00 | 0.37 | - | - | - | - | - |
| addr\_state | атрибутивный | 0.00 | 0.02 | - | - | - | - | - |
| dti | вещественный | 0.00 | 1.70 | 18.041 | 17.630 | 8.023 | 39.990 | 0.000 |
| delinq\_2yrs | целочисленный | 0.00 | 0.01 | 0.345 | 0.000 | 0.898 | 22.000 | 0.000 |
| earliest\_cr\_line | атрибутивный | 0.00 | 0.27 | - | - | - | - | - |
| inq\_last\_6mths | целочисленный | 0.00 | 0.00 | 0.756 | 0.000 | 1.034 | 6.000 | 0.000 |
| mths\_since\_last\_delinq | вещественный | 49.18 | 0.05 | 33.398 | 30.000 | 21.780 | 188.000 | 0.000 |
| mths\_since\_last\_record | вещественный | 82.38 | 0.05 | 70.711 | 69.000 | 28.463 | 121.000 | 0.000 |
| open\_acc | целочисленный | 0.00 | 0.03 | 11.668 | 11.000 | 5.268 | 84.000 | 0.000 |
| pub\_rec | целочисленный | 0.00 | 0.01 | 0.222 | 0.000 | 0.605 | 63.000 | 0.000 |
| revol\_bal | целочисленный | 0.00 | 20.71 | 16508.089 | 11686.000 | 21462.891 | 2560703.000 | 0.000 |
| revol\_util | вещественный | 0.05 | 0.49 | 0.557 | 0.563 | 0.231 | 8.923 | 0.000 |
| total\_acc | целочисленный | 0.00 | 0.05 | 26.011 | 24.000 | 11.896 | 156.000 | 2.000 |
| initial\_list\_status | атрибутивный | 0.00 | 0.00 | - | - | - | - | - |
| collections\_12\_mths\_ex\_med | целочисленный | 0.00 | 0.00 | 0.015 | 0.000 | 0.143 | 20.000 | 0.000 |
| mths\_since\_last\_major\_derog | вещественный | 71.79 | 0.07 | 43.370 | 43.000 | 21.962 | 188.000 | 0.000 |
| policy\_code | целочисленный | 0.00 | 0.00 | 1.000 | 1.000 | 0.000 | 1.000 | 1.000 |
| application\_type | атрибутивный | 0.00 | 0.00 | - | - | - | - | - |
| acc\_now\_delinq | целочисленный | 0.00 | 0.00 | 0.006 | 0.000 | 0.081 | 4.000 | 0.000 |
| tot\_coll\_amt | целочисленный | 0.00 | 2.45 | 270.302 | 0.000 | 18952.919 | 9152545.000 | 0.000 |
| tot\_cur\_bal | целочисленный | 0.00 | 67.17 | 139802.267 | 82027.000 | 153022.613 | 4026405.000 | 0.000 |
| acc\_open\_past\_24mths | целочисленный | 0.00 | 0.02 | 4.405 | 4.000 | 2.865 | 53.000 | 0.000 |
| avg\_cur\_bal | вещественный | 0.00 | 20.20 | 13413.435 | 7533.000 | 16016.231 | 497484.000 | 0.000 |
| bc\_open\_to\_buy | вещественный | 1.04 | 16.33 | 8461.743 | 3625.000 | 13365.346 | 260250.000 | 0.000 |
| bc\_util | вещественный | 1.11 | 0.53 | 64.659 | 68.900 | 26.433 | 255.200 | 0.000 |
| chargeoff\_within\_12\_mths | целочисленный | 0.00 | 0.00 | 0.011 | 0.000 | 0.117 | 7.000 | 0.000 |
| delinq\_amnt | целочисленный | 0.00 | 0.23 | 10.200 | 0.000 | 587.501 | 70076.000 | 0.000 |
| mo\_sin\_old\_il\_acct | вещественный | 3.04 | 0.19 | 128.533 | 131.000 | 51.280 | 561.000 | 0.000 |
| mo\_sin\_old\_rev\_tl\_op | целочисленный | 0.00 | 0.29 | 185.749 | 168.000 | 93.041 | 842.000 | 3.000 |
| mo\_sin\_rcnt\_rev\_tl\_op | целочисленный | 0.00 | 0.09 | 13.070 | 8.000 | 16.112 | 372.000 | 0.000 |
| mo\_sin\_rcnt\_tl | целочисленный | 0.00 | 0.06 | 7.997 | 6.000 | 8.739 | 226.000 | 0.000 |
| mort\_acc | целочисленный | 0.00 | 0.01 | 1.851 | 1.000 | 2.161 | 37.000 | 0.000 |
| mths\_since\_recent\_bc | вещественный | 0.95 | 0.16 | 24.445 | 14.000 | 30.296 | 616.000 | 0.000 |
| mths\_since\_recent\_bc\_dlq | вещественный | 73.57 | 0.06 | 39.631 | 39.000 | 22.573 | 170.000 | 0.000 |
| mths\_since\_recent\_inq | вещественный | 9.21 | 0.01 | 6.920 | 5.000 | 5.931 | 25.000 | 0.000 |
| mths\_since\_recent\_revol\_delinq | вещественный | 64.03 | 0.06 | 35.459 | 32.000 | 22.299 | 180.000 | 0.000 |
| num\_accts\_ever\_120\_pd | целочисленный | 0.00 | 0.01 | 0.503 | 0.000 | 1.264 | 33.000 | 0.000 |
| num\_actv\_bc\_tl | целочисленный | 0.00 | 0.01 | 3.687 | 3.000 | 2.156 | 26.000 | 0.000 |
| num\_actv\_rev\_tl | целочисленный | 0.00 | 0.02 | 5.805 | 5.000 | 3.141 | 38.000 | 0.000 |
| num\_bc\_sats | целочисленный | 0.00 | 0.02 | 4.648 | 4.000 | 2.724 | 35.000 | 0.000 |
| num\_bc\_tl | целочисленный | 0.00 | 0.02 | 8.544 | 8.000 | 4.822 | 61.000 | 0.000 |
| num\_il\_tl | целочисленный | 0.00 | 0.04 | 8.573 | 7.000 | 7.303 | 150.000 | 0.000 |
| num\_op\_rev\_tl | целочисленный | 0.00 | 0.02 | 8.277 | 7.000 | 4.319 | 62.000 | 0.000 |
| num\_rev\_accts | целочисленный | 0.00 | 0.03 | 15.299 | 14.000 | 8.055 | 105.000 | 2.000 |
| num\_rev\_tl\_bal\_gt\_0 | целочисленный | 0.00 | 0.02 | 5.770 | 5.000 | 3.124 | 38.000 | 0.000 |
| num\_sats | целочисленный | 0.00 | 0.03 | 11.619 | 11.000 | 5.266 | 84.000 | 0.000 |
| num\_tl\_120dpd\_2m | вещественный | 3.34 | 0.00 | 0.001 | 0.000 | 0.032 | 2.000 | 0.000 |
| num\_tl\_30dpd | целочисленный | 0.00 | 0.00 | 0.004 | 0.000 | 0.065 | 4.000 | 0.000 |
| num\_tl\_90g\_dpd\_24m | целочисленный | 0.00 | 0.01 | 0.095 | 0.000 | 0.491 | 22.000 | 0.000 |
| num\_tl\_op\_past\_12m | целочисленный | 0.00 | 0.01 | 2.007 | 2.000 | 1.605 | 26.000 | 0.000 |
| pct\_tl\_nvr\_dlq | вещественный | 0.00 | 0.20 | 94.240 | 97.600 | 8.463 | 100.000 | 16.700 |
| percent\_bc\_gt\_75 | вещественный | 1.09 | 0.07 | 50.770 | 50.000 | 34.916 | 100.000 | 0.000 |
| pub\_rec\_bankruptcies | целочисленный | 0.00 | 0.00 | 0.135 | 0.000 | 0.376 | 12.000 | 0.000 |
| tax\_liens | целочисленный | 0.00 | 0.01 | 0.056 | 0.000 | 0.409 | 63.000 | 0.000 |
| tot\_hi\_cred\_lim | целочисленный | 0.00 | 67.03 | 170248.977 | 111514.000 | 171996.126 | 9999999.000 | 0.000 |
| total\_bal\_ex\_mort | целочисленный | 0.00 | 41.72 | 48389.433 | 36685.000 | 45966.066 | 2688920.000 | 0.000 |
| total\_bc\_limit | целочисленный | 0.00 | 1.21 | 20031.415 | 13800.000 | 20209.994 | 1090700.000 | 0.000 |

Удалим признаки, доля пропусков в которых более 70% и атрибутивные признаки, в которых количество уникальных элементов более 300, ибо если взять менее строгое правило, то это приведет к существенному увеличению объема данных и сложностям в процессе обучения нейронной сети.

Отделим интересующий признак grade от общей совокупности. Также удалим признак sub\_grade, который является уточнением к признаку grade, и признак loan\_status, имеющий прямую связь с целевым признаком. Таким образом, мы получили матрицу свободных переменных и вектор зависимых. Приведем все значения признаков в числовой формат. Так, все атрибутивные признаки переводятся в булево пространство, причем отсутствие данных это отдельное значение признака. Представляем целевой признак grade, который принимает следующие значения: A, B, C, D, E, F, в числовом формате . Заполним пропуски для числовых значений средним значением признака.

Важным условием обучение сети является стандартизация распределения значений признака. То есть признак должен принимать значения в небольшой окрестности нуля. Для этого из каждого старого значения вычтем математическое ожидание признака и поделим на стандартное отклонение.

Так мы получили матрицу со свободными признаками и вектор целевых признаков.