Project Objective
Telecommunication companies experience customer churn when customers leave to a competitor's service. A key business problem is identifying customers that are likely to churn to retain them.

Deep learning will be utilised to identify customers who will churn based on their characteristics so that they can be retained through special offers. Therefore a successful model will identify these customers at a high rate, while minimising the number of non-churn customers who receive the special (lower margin) offers.

Data Set
In practice this application would be achieved through the use of a company's internal customer meta data. It might also include usage information or other profiles developed in house. For the current project a dataset has been obtained from Kaggle [1].

The data columns are given in table A1 in the appendix. The total data size is 7,043 samples with 1,869 positive, representing 26.5% of the data. This is generally considered a split low for binary classification using DNN, so a data stratification will be performed.

Model
A multi-layer DNN model has been selected for the project. Initial testing will use the 'Relu' activation function for the input and hidden layers, with 'sigmoid' utilised for the output layer.

The baseline structure was an input layer of 8 neurons and an 8 neuron hidden layer. The output layer was 1 neuron as it is a binary classification task. A model is presented in figure A2.

Table 1. Key starting model parameters.

Compile Parameters		Fit Parameters	
Loss	binary_crossentropy	Epochs	20
Optimzer	adam	Validation split	0.25

The compile stage used the accuracy metric, as well as the FlaseNegatives, TrueNegatives, FalsePositives and TruePositives. These extra metrics were used to calculate the performance.

Experiments
Experiments were evaluated against two key sets of metrics: the ratio of true positives to total positives in validation set (i.e. accuracy at predicting positives), total accuracy and loss.

1) Data size
The first test was to select an idea dataset size for the model. A correlation was performed to identify attributes with a strong relationship to the churn status. Results are given in table A2.

Further tests were then performed progressively removing features with low correlation to reduce noise and model size. Tests were performed with big (64, 32, 1 neuron) and small (8, 8, 1 neuron) network sizes to avoid bias due to smaller and larger datasets.

A strong correlation between contract type "month-to-month" and churn was found. This was used to stratify data, with a test performed using only this data. However the more balanced dataset performed worse than the base line test in terms of true positive rate.

It was concluded that the smallest data sets dropped performance, perhaps due to loss of data features or the smaller volume of data. Consequently the datasets were expanded with a one hot on two columns with high correlation items (contract type and internet service). The best result was for the original 9 column dataset plus the one hot on the highest correlation columns. Select results are shown below, with full results in table A3.

Parameter	Data Notes	Loss, Accuracy	True Positives Ratio
All data	Big network	0.94, 0.79	0.357
Top 9 Columns	Big network	0.41, 0.81	0.582
Top 5 Columns	Small network	0.41, 0.81	0.568
Stratified Data	4 columns	0.61, 0.66	0.548
9 Columns with OH	One hot	0.42, 0.81	0.619

2) Model Parameters
The second experiment tested the sensitivity of the model to changes in different model parameters. Select results are given in the table below, with full results in table A4.

Parameter	Test Range	Loss, Accuracy	True Positives Ratio
Base line model		0.41, 0.81	0.582
Epoch	10	0.41, 0.81	0.609
	50	0.41, 0.81	0.625
Hidden layer count	0	0.41, 0.81	0.579
	2	0.42, 0.81	0.619
Input & hidden layer activation	Softmax	0.43, 0.8	0.668

It was considered that for the business application it was more economic to select a model with more false positives than false negatives (i.e. cost of missing a potential churn customer is higher than offering a deal to a customer who wasn't going to churn). Therefore an adjusted softmax model was adopted.

3) Optimising model
The adjusted softmax model was then optimised for over and underfitting using dropout, regularisers and epoch length. It was found that dropout and regularisers tended to skew the results and therefore reduced model accuracy, so were not implemented in the final model. Full results are given in table A5, with final model parameters given in table A6.

Conclusion
The model was successfully implemented for the business problem. Further study should focus on collecting extra data to provide more variables with higher correlation to churn. Alternative models could be implemented which consider the probability that a customer will churn and offer a graded retention offer.

Appendix

Table A1. Input data.

Tag	Description	Tag	Description
customerID	Unique identifier.	DeviceProtection	Add on service.
gender	Customer gender.	TechSupport	Add on service.
SeniorCitizen	Age related attribute.	StreamingTV	Add on service.
Partner	Social status attribute.	StreamingMovies	Add on service.
Dependents	Social status attribute.	Contract	Contract type.
tenure	Length with company.	PaperlessBilling	Payment type attribute.
PhoneService	Holds a phone service.	PaymentMethod	Payment method.
MultipleLines	Multiple phone lines.	MonthlyCharges	Amount paid.
InternetService	Holds an internet service and type.	TotalCharges	Total amount paid.
OnlineSecurity	Add on service.	Churn	Customer status.
OnlineBackup	Add on service.		

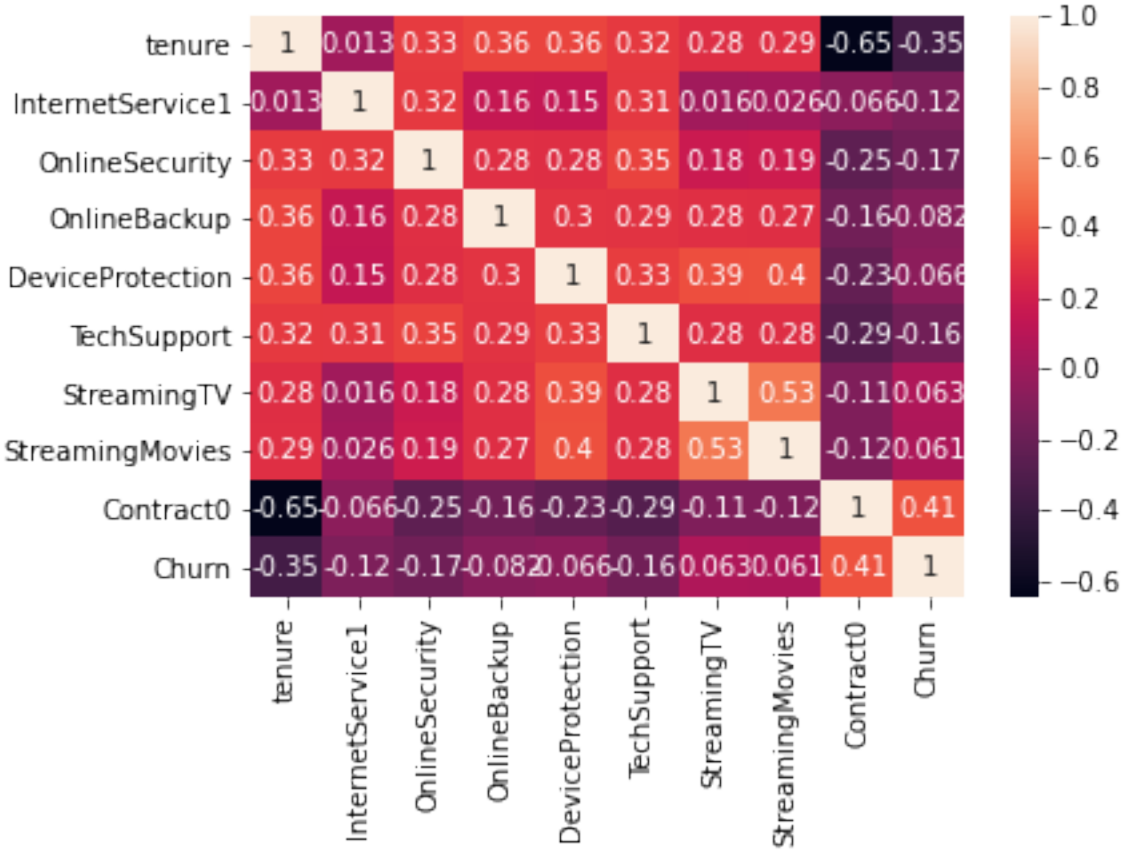


Figure A1. Correlation heat map for final dataset.

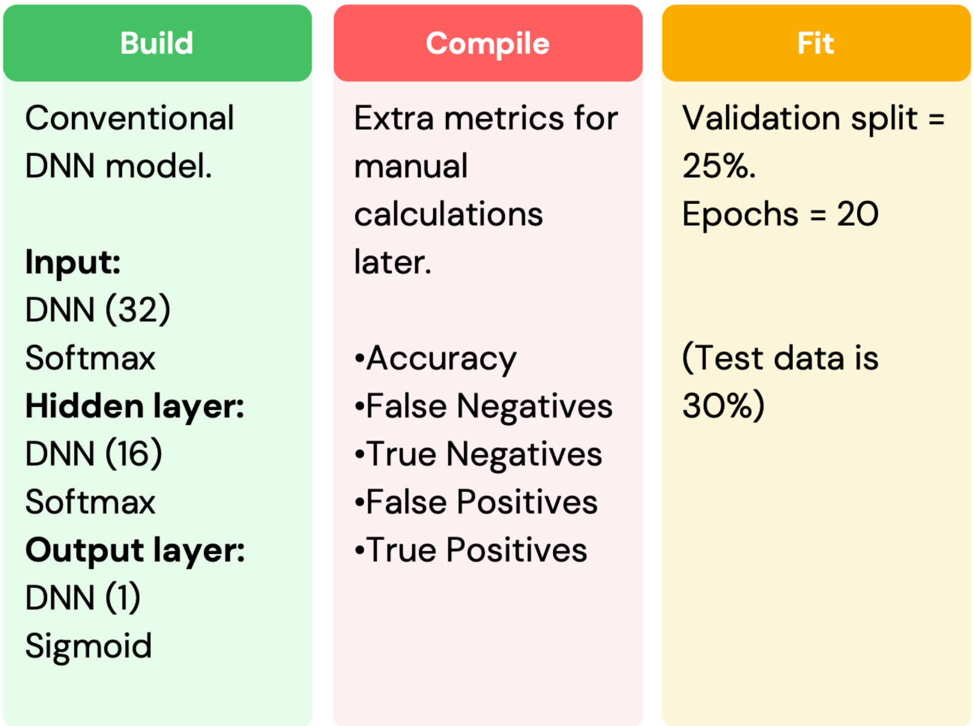


Figure A2. Diagram of network.

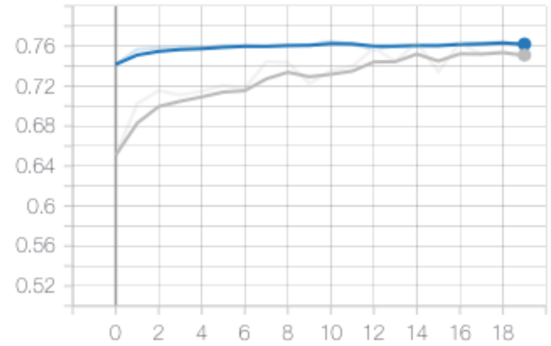
Table A2. Correlation to Churn column in original data file.

Column	Correlation
gender	0.01
SeniorCitizen	0.15
Partner	-0.15
Dependents	-0.16
tenure	-0.35
PhoneService	0.01
MultipleLines	0.04
InternetService	0.32
OnlineSecurity	-0.17
OnlineBackup	-0.08
DeviceProtection	-0.07
TechSupport	-0.16
StreamingTV	0.06
StreamingMovies	0.06
Contract	-0.40
PaperlessBilling	0.19
PaymentMethod	-0.26
MonthlyCharges	0.19
Churn	1.00

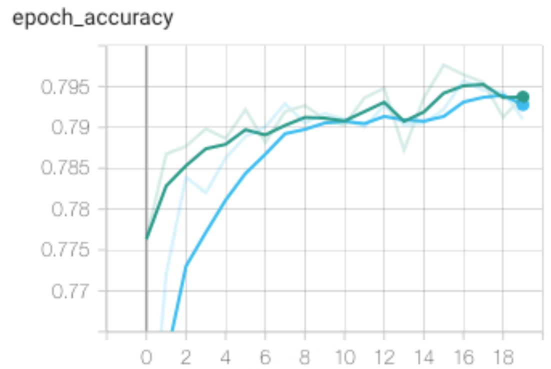
Table A3. Full results for dataset optimisation (test results.

Parameter	Data Notes	Loss, Accuracy	True Positives Ratio
All data	Big network	0.94, 0.79	0.357
Top 9 Columns	Big network	0.41, 0.81	0.582
Top 5 Columns	Small network	0.41, 0.81	0.568
Stratified Data	Small network 4 columns	0.61, 0.66	0.548
5 Columns with one hot	One hot on Internet and Contract	0.41, 0.81	0.547
9 Columns with one hot	One hot on Internet and Contract	0.42, 0.81	0.619

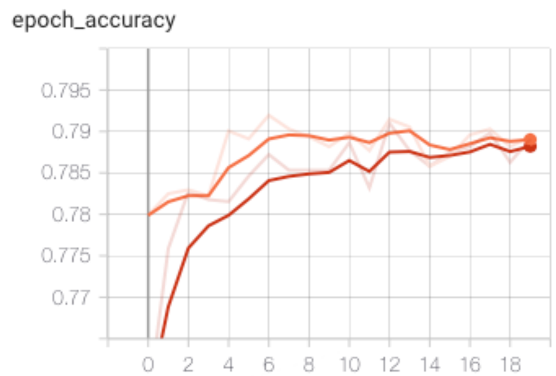
Table A3a. Tensorboard results for experiment – training data.



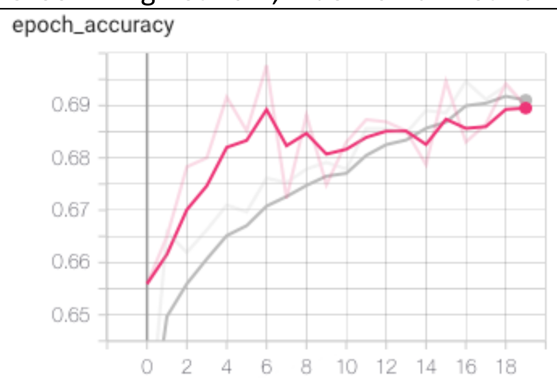
All data.
Blue = Big network, Grey = Small network



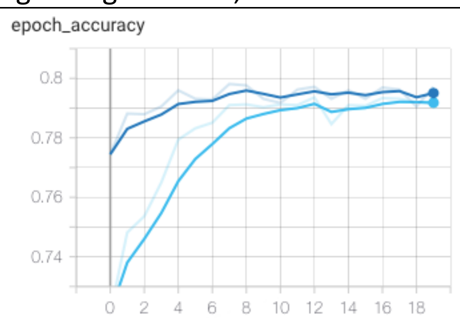
Top 9 columns.
Green = Big network, Blue = Small network



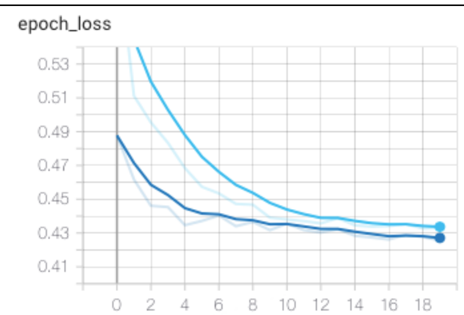
Top 5 columns.
Orange = Big network, Red = Small network



Stratified data.
Pink = Small network, grey = Big nework



9 column with one hot
Dark = Small network, light = big network

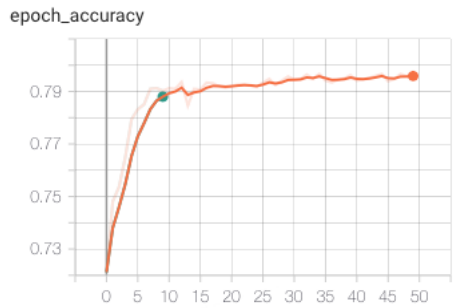


9 column with one hot
Dark = Small network, light = big network

Table A4. Full results for model optimisation.

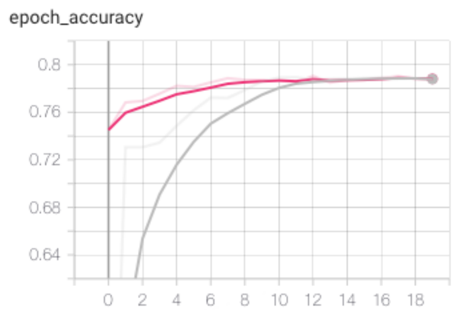
Parameter	Test Range	Loss, Accuracy	True Positives Ratio
Base line model		0.41, 0.81	0.582
Epoch	10	0.41, 0.81	0.609
	50	0.41, 0.81	0.625
Output activation	Softmax	11.21, 0.26	1
Optimizer	sgd	0.5, 0.76	0.212
Hidden layer count	0	0.41, 0.81	0.579
	2	0.42, 0.81	0.619
Hidden layer size	4	0.41, 0.81	0.573
	16	0.41, 0.81	0.568
Input & hidden layer activation	Softmax	0.43, 0.8	0.668
	Relu(1) + Softmax(2)	0.43, 0.81	0.560

Table A4a. Tensorboard results for experiment – training data.

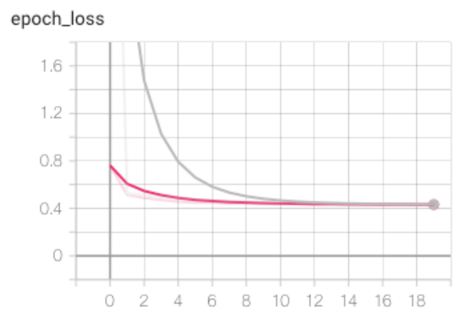


Orange = Long epoch, Grey = Short epoch

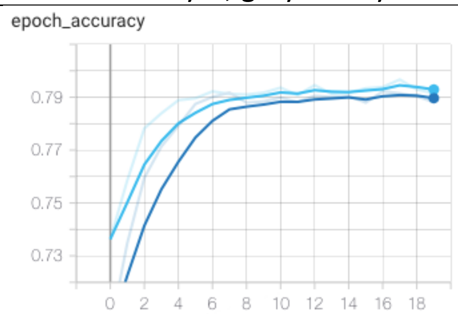
Orange = Long epoch, Grey = Short epoch



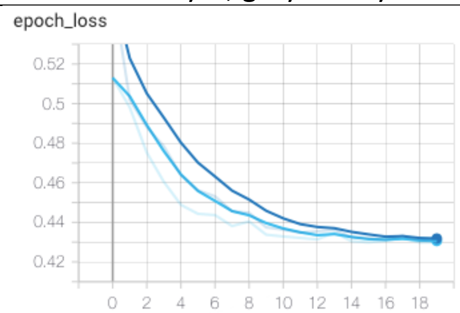
Pink = 2 layer, grey = 4 layer



Pink = 2 layer, grey = 4 layer



Dark = bigger layer, light = smaller layer

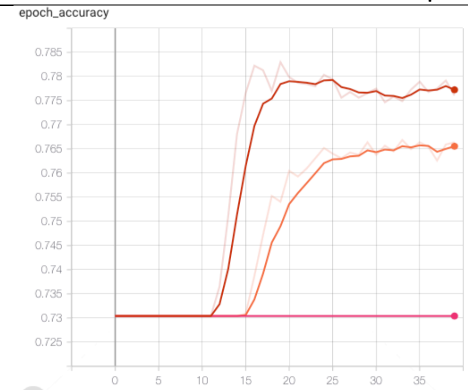


Dark = bigger layer, light = smaller layer

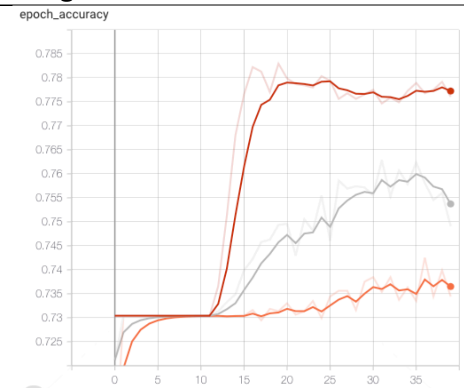
Table A5. Full results for final model optimisation.

Parameter	Test Range	Loss, Accuracy	True Positives Ratio
Base line model		0.45, 0.78	0.617
Regularizer	L1 = 0.001	0.49, 0.74	0
	L2 = 0.001	0.46, 0.78	0.397
	L2 = 0.01	0.57, 0.74	0
Dropout	0.5	0.47, 0.74	0
	0.2	0.45, 0.77	0.491

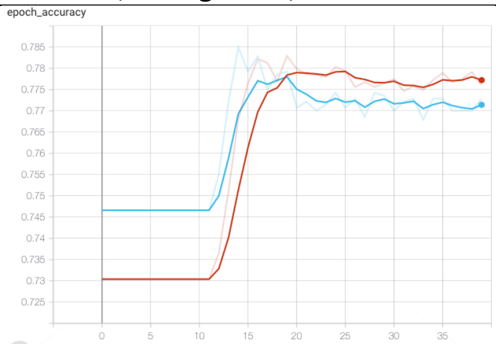
Table A5a. Tensorboard results for experiment – training data.



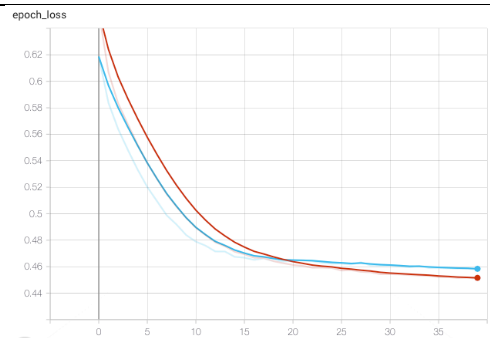
Pink = L1, Orange = L2, Red = Baseline



D=0.5 Orange, D=0.2 Grey, Red = Baseline



Orange = training, Blue = validation

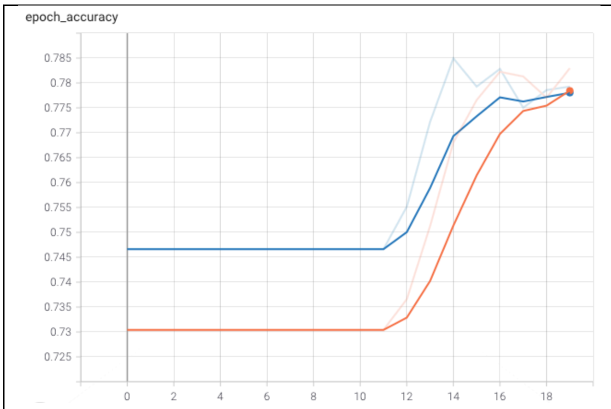


Orange = training, Blue = validation

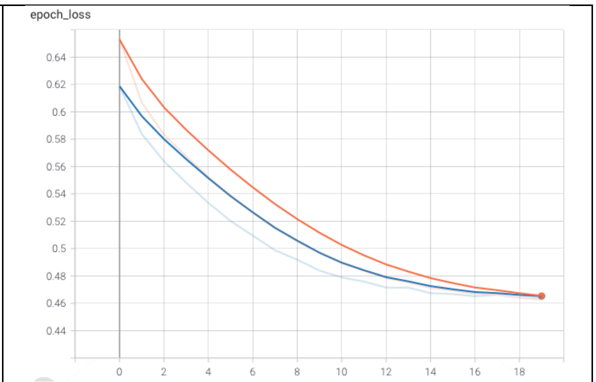
Table A6. Final model parameters.

Compile Parameters		Fit Parameters	
Loss	binary_crossentropy	Epochs	20
Optimzer	adam	Validation split	0.25
Metrics	"accuracy", tf.keras.metrics.FalseNegatives(), tf.keras.metrics.TrueNegatives(), tf.keras.metrics.FalsePositives(), tf.keras.metrics.TruePositives()		
Input and Hidden Layers		**Output Layer**	
Neurons	32, 16	Neurons	1
Activation	Softmax	Activation	Sigmoid

Table A6a. Tensorboard results for final model.



Orange = training, Blue = validation



Orange = training, Blue = validation

References

[1] *Telco Customer Churn*, Kaggle, Blastchar, Accessed 2020/12/18
https://www.kaggle.com/blastchar/telco-customer-churn