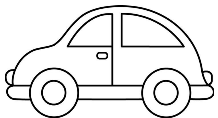
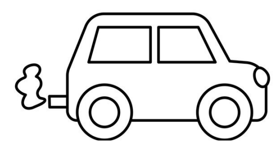


Random Forest: Traffic Accidents Data Exploration

Brendan Callender

Martin Hsu

Kyle Lew

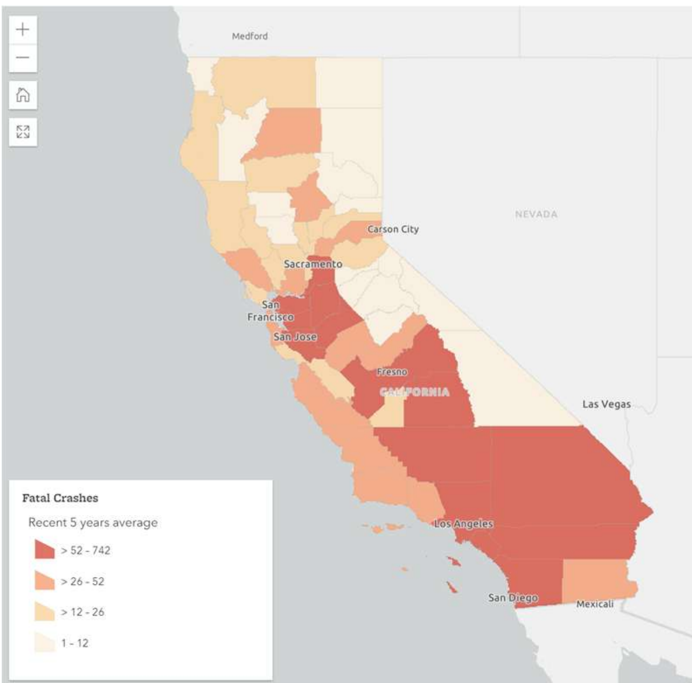


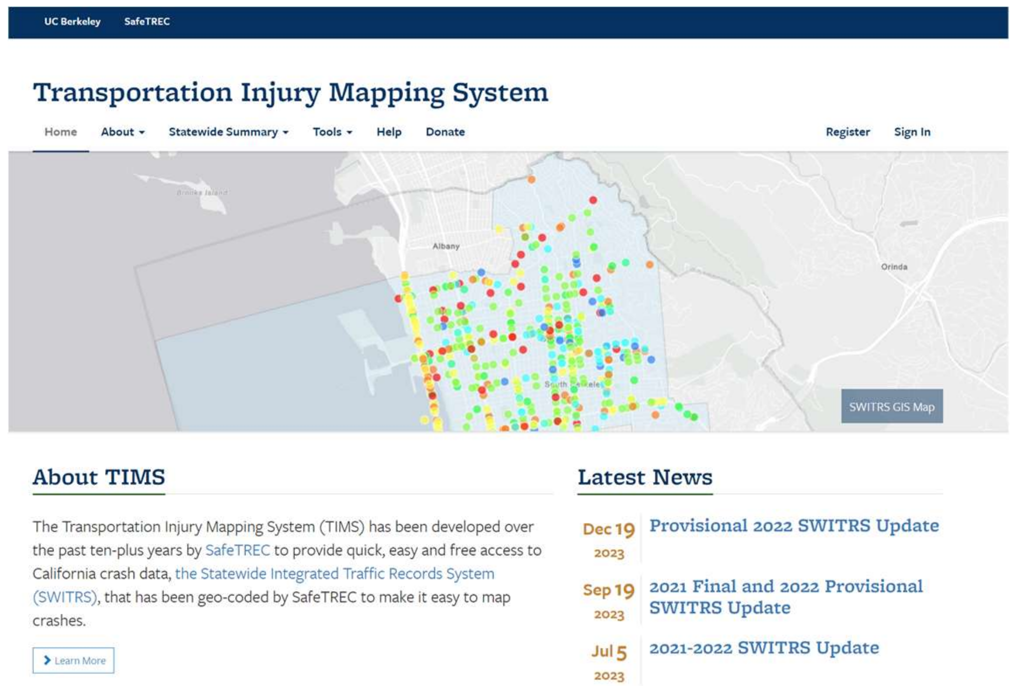
Goal

- Machine Learning: Random Forest Classification
- Data: Traffic Accident Data
- Questions: Based on traffic data, can we predict…
  - Collision severity?
  - Weather?
  - Collision type?

Data Description

- California Highway Patrol data
  - Provided by UC Berkeley





Dataset Example

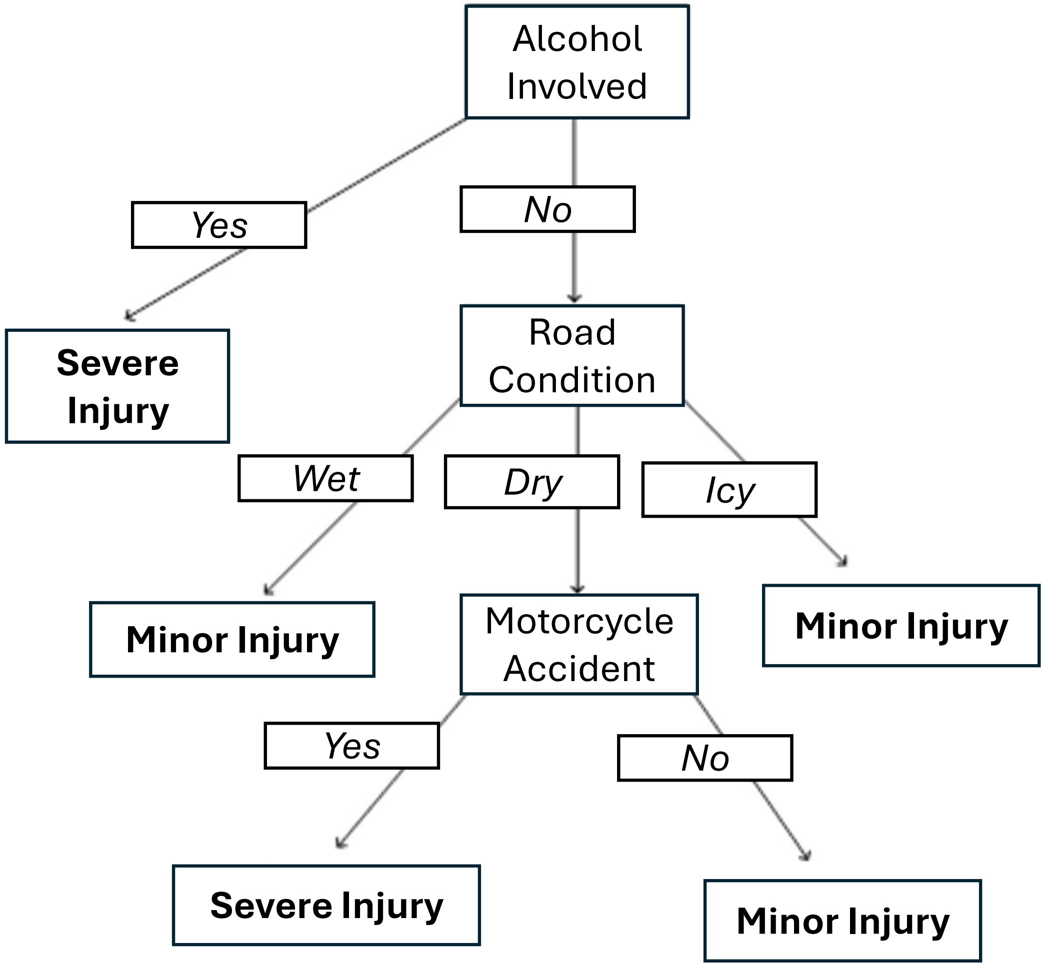
CASE_ID	ACCIDENT	PROC_DATE	JURIS	COLLISION_DATE	COLLISION	OFFICER_I	REPORTIN	DAY_OF_W	CHP_SHIF	POPULATI	CNTY_CITY	SPECIAL_C	BEAT_TYPE	CHP_BEAT	CITY_DIVIS	CHP_BEAT	BEAT_NUM	PRIMARY_I	SECONDA	DISTANCE	DIRECTION	INTERSECT	WEA
90284372	2016	10/5/2016	9252	9/22/2016	1545	17869		4	2	9	3400	0	1	3		2	191	SR-99 S/B	MACK RD	2640	S	N	A
90393042	2017	2/14/2017	9252	2/11/2017	915	19231		6	1	9	3400	0	2	4		1	72	MARTIN LU	43RD AVE	0		Y	A
90625118	2017	12/22/2017	9260	12/18/2017	1028	19625		1	1	9	3400	0	2	4		2	307	GREENBAC	MAIN AVE	2	N	N	A
90692594	2018	3/29/2018	9250	3/19/2018	733	17445		1	1	9	3400	0	2	4		1	12	ANTELOPE	ANTELOPE	122	E	N	A
91020587	2019	6/28/2019	9252	6/20/2019	1345	16070		4	1	7	3404	0	1	1		2	108	I-5 NB	RICHARDS	682	S	N	A
91271386	2020	12/11/2020	9252	7/10/2020	1826	21550		5	2	9	3400	0	3	5		2	74	MORNING	HOPYARD	0		Y	A
91517336	2021	7/12/2021	9250	7/6/2021	1716	17568		2	2	9	3400	0	1	1		2	184	MADISON	MADISON	300	W	N	A
8802969	2019	2/25/2019	3404	2/17/2019	1626	8162	2	7	5	7	3404	0	0	0		0	00C	EL CAMINO	PRINCETO	580	E	N	A
9117330	2020	9/15/2020	3404	6/13/2020	1737	1053	P2	6	5	7	3404	0	0	0		0	02A	ALDER ST	NORTH AV	93	S	N	A
6617204	2014	12/21/2015	9250	7/31/2014	400	18785		4	3	9	3400	0	2	4		1	61	RIO LINDA	ELKHORN	1108	S	N	A
6308127	2013	5/30/2014	9252	12/9/2013	1750	18159		1	2	7	3404	0	1	3		2	192	RT 99	BROADWA	1056	S	N	A
6038834	2013	2/12/2014	3400	3/5/2013	2159	03CSO	RC	2	5	5	3490	0	0	0		0	5	MATHER F	ROCKINGH	0		Y	A
90308142	2016	11/1/2016	9250	8/19/2016	2019	18374		5	2	9	3400	0	2	4		1	21	AUBURN B	MADISON	260	S	N	A
90365967	2016	6/2/2017	9250	12/23/2016	1805	16957		5	2	7	3404	0	1	3		2	152	SR-51 S/B	SR-160	750	N	N	B
90433720	2017	4/13/2017	9252	3/26/2017	1620	19465		7	2	9	3400	0	2	4		1	72	FLORIN RC	FRANKLIN	250	E	N	B
90812419	2018	9/13/2018	9250	9/7/2018	1245	15423		5	1	9	3400	0	2	4		1	24	MADISON	DATE AVEN	40	W	N	A
90957617	2019	3/29/2019	9250	3/18/2019	1735	20602		1	2	9	3400	0	3	5		2	40	BELL AVE	IRMA WAY	100	S	N	A
91425584	2021	3/11/2021	9252	3/5/2021	1915	20783		5	2	9	3400	0	2	4		2	73	STOCKTON	ORANGE A	639	N	N	A
6032185	2013	2/10/2014	340H	3/26/2013	736	10111	5504	2	5	5	3450	0	0	0		0	5	ELKMONT	IRON ROC	776	W	N	A
91561067	2021	9/2/2021	9250	8/26/2021	745	18116		4	1	9	3400	0	3	5		2	10	PALM AVE	HILLSDALE	219	W	N	A
9242026	2021	4/13/2021	340H	3/24/2021	1639	10799	4104	3	5	5	3450	0	0	0		0	4	ELK GROVI	BACKER R	793	W	N	A
7196447	2016	3/15/2016	3404	3/1/2016	1346	8169		2	5	7	3404	0	0	0		0		20TH ST	P ST	0		Y	A
8051367	2016	6/2/2016	340H	5/27/2016	1806	10573	5401	5	5	5	3450	0	0	0		0	5	ELK GROVI	EMERALD	85	E	N	A
7018809	2014	6/28/2016	9250	11/1/2014	126	17317		6	3	5	3496	0	1	1		2	183	RT 80 WB	WEST OF F	4752		N	B
5636445	2012	2/24/2014	9252	4/30/2012	1515	17917		1	2	7	3404	0	1	2		2	151	RT 50	STOCKTON	1584	E	N	A
90132976	2016	3/9/2016	9252	3/2/2016	1440	18423		3	2	9	3400	0	2	4		1	71	FRUITRIDG	MENDOCII	150	E	N	A
90728175	2018	5/17/2018	9250	5/11/2018	2028	19572		5	2	9	3400	0	3	5		2	20	GREENBAC	GARFIELD	760	W	N	A
90763260	2018	7/5/2018	9252	6/23/2018	2222	20520		6	3	7	3404	0	1	3		2	193	SR-51 S/B	MCKINLEY	370	N	N	A
90952573	2019	3/25/2019	9260	3/18/2019	2105	18370		1	2	9	3400	0	3	5		2	1	KIEFER BL	BRADSHAV	0		Y	A
91330385	2020	10/23/2020	9252	10/13/2020	1800	18730		2	2	7	3404	0	1	1		2	107	I-5 N/B	SUTTERVIL	1200	S	N	A
9204650	2020	2/13/2021	340H	12/26/2020	124	10754	5404	6	5	5	3450	0	0	0		0	5	ELK GROVI	E STOCKT	0		Y	C
9242987	2021	4/15/2021	3404	3/15/2021	1502	8107	P2	1	5	7	3404	0	0	0		0	02A	NORWOO	JESSIE AV	0		Y	A
90006432	2015	10/19/2015	9250	7/22/2015	905	14845		3	1	9	3400	0	1	1		2	183	I-80 W/B T	MADISON	5	E	N	A
8939395	2019	11/21/2019	3404	10/23/2019	650	304	6	3	5	7	3404	0	0	0		0	06C	14TH AV	DON MERI	0		Y	A
5493995	2012	7/13/2013	3496	2/20/2012	943	192	3	1	5	5	3496	0	0	0		0	3	GREENBAC	LONGFOR	75	W	N	A

Variables Used

- **All variables**:
  - Day of Week
  - Intersection?
  - Weather
  - State Highway?
  - Tow Away?
  - Collision Severity
  - Party Count
  - Primary Collision Factor
  - Hit and Run
  - Type of Collision
  - Road Surface
  - Road Condition
  - Lighting
  - Pedestrian Accident?
  - Bicycle Accident?
  - Motorcycle Accident?
  - Truck Accident?
  - Alcohol Involved?
- Trying to predict
  - **Collision Severity**
  - **Weather**
  - **Collision Type**
- Use other variables to predict

Decision Trees

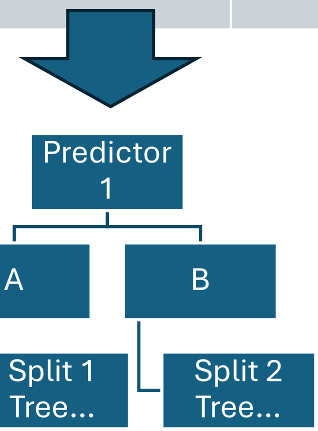
- Tree to classify observations according to categorical attributes
- Tree Structure
  - Internal Node = attribute
  - Edges = specific values of an attribute
  - Leaf Node = classification decision



Implementing Decision Trees(C4.5)

- Recursive algorithm
- Choose tree splits based on entropy and information gain
  - Entropy – how similar the class attributes are in a group
  - Information gain – do attributes within splits become more similar after splitting the data?

Predictor 1	Predictor 2	Class Attribute
A	C	T
A	C	T
A	D	T
B	D	F
B	D	T
B	C	F
B	C	F



C4.5 Code

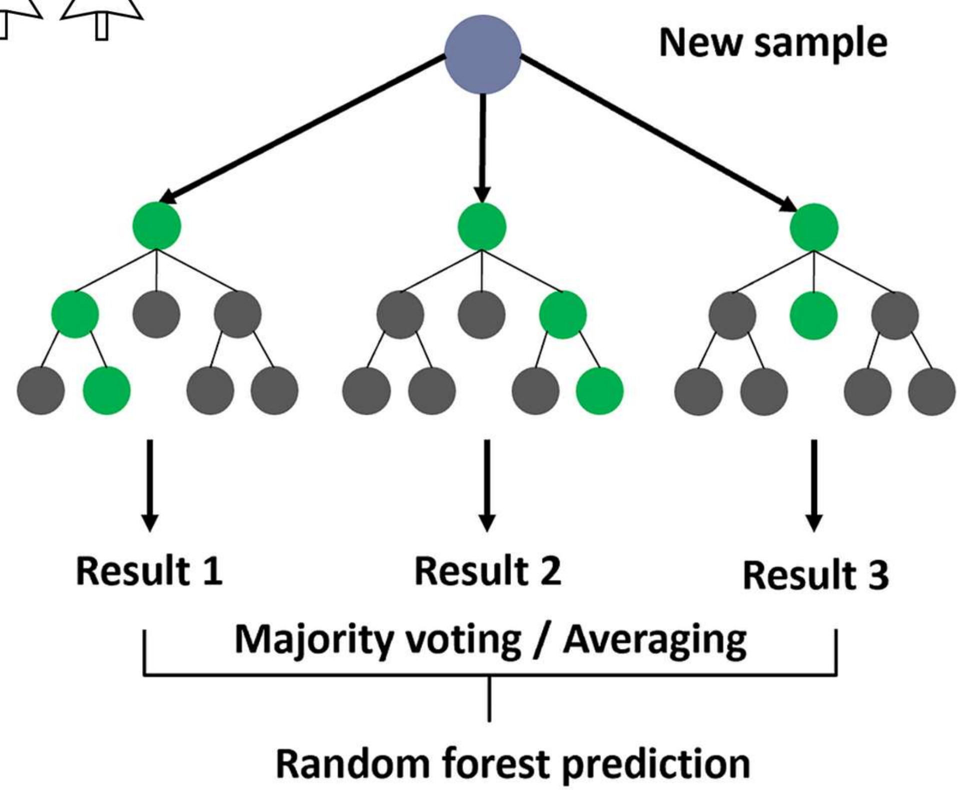
- Base Cases
- Recursive Step

```scala
def C45(D: RDD[Array[String]], C: RDD[String], A: Array[String], domain: Map[String, Array[String]],
        threshold: Double): Tree = {
  var T = new Tree
  if (C.distinct.count == 1) {
    T.setRootNode(Leaf(C.take( num = 1)(0), 1.0))
  }
  else if (A.length == 0) {
    var label = findMostFreqLabel(C)
    T.setRootNode(Leaf(label._1, label._2))
  }
```

```scala
    else {
      var branch = Branch(A_g.get)
      T.setRootNode(branch)

      var A_g_idx = A.indexOf(A_g.get)
      var A_g_vals = domain(A_g.get)

      def makeBranch(v: String): Unit = {
        val cZipA = C.zip(D.map(x => x(A_g_idx)))

        var D_v = D.filter(x => x(A_g_idx) == v).map(x => removeIndex(x, A_g_idx)).persist
        var C_v = cZipA.filter(x => x._2 == v).keys.persist

        if (D_v.count != 0) {
          val A_v = A.filter(x => x != A_g.get)
          val T_v = C45(D_v, C_v, A_v, domain, threshold)

          D_v = D_v.unpersist()
          C_v = C_v.unpersist()

          val newEdge = Edge(v)
          newEdge.setNode(T_v.rootNode)

          branch.addToEdges(newEdge)
        }
        else {
          val label = findMostFreqLabel(C)
          val newEdge = Edge(v)
          newEdge.setNode(Leaf(label._1, label._2))

          branch.addToEdges(newEdge)
        }
      }
      A_g_vals.foreach(v => makeBranch(v))
    }
  }
  T
}
```

Random Forest

- Construct many Decision Trees
  - Voting to determine final predicted class
- Created using a subset of the data
  - Sample observations w replacement (rows)
    - Simple random sampling
    - Stratified sampling
  - Small subset of predictors (columns)



Random Forest Code

```scala
mshsu
def fitTrees(D: RDD[Array[String]], C: RDD[String], A: Array[String],
             nTrees: Int, nAttrs: Int, nSample: Int,
             threshold: Double, stratified: Boolean): Array[Tree] = {
  val enumTrees = Array.range(1, nTrees)
  println("Bootstrapping...")
  val samples = enumTrees.map(x => bootstrap(D, C, A, nAttrs, nSample, stratified))
  println("Training random forest...")
  val domain = InduceC45.getAttrDomain(D, A)
  samples.map({ case (d, c, a) => InduceC45.C45(d, c, a, domain, threshold) })
}

mshsu
def classify(row: Array[String], A: Array[String], trees: Array[Tree]): String = {
  val votes = trees.map(t => Classify.classify(row, A, t))
  val decision = votes.groupBy(identity).mapValues(_.length).maxBy(_._2)._1
  decision
}
```

Model Evaluation

- **Accuracy** - # correct overall
- For each category:
  - **Precision** – Out of those predicted to be A, how many observed to be A
  - **Recall** – Out of those observed to be A, how many predicted to be A
  - **F1 Score** – harmonic mean of precision and recall

- Random Forest Parameters:
  - 50 Trees
  - 50 observations
  - 3 variables used
  - 0.1 Information Gain Threshold

Results: Collision Severity

- "Possible injury or complaint of pain" always predicted
    - Overrepresented

Predicting Collision Severity

Accuracy = 0.64

Table 1: Prediction Results for Collision Severity

Collision Severity	Precision	Recall	F1-Score
Fatal Injury	NA	0	NA
Suspected serious injury of severe injury	NA	0	NA
Suspected minor injury of visible injury	NA	0	NA
Possible injury of complain of pain	0.64	1	0.78

Results: Weather

- “Clear” always predicted
  - Overrepresented

Predicting Weather

Accuracy = 0.89

Table 2: Prediction Results for Weather

Weather	Precision	Recall	F1-Score
Not Stated	NA	0	NA
Clear	0.89	1	0.94
Cloudy	NA	0	NA
Raining	NA	0	NA
Snowing	NA	0	NA
Fog	NA	0	NA
Other	NA	0	NA
Wind	NA	0	NA

Results: Collision Type

- Better representation

Predicting Collision Type

Accuracy = 0.48

Table 3: Prediction Results for Collision Type

Collision Type	Precision	Recall	F1-Score
Not Stated	NA	0	NA
Head-On	NA	0	NA
Sideswipe	NA	0	NA
Rear End	0.44	0.90	0.59
Broadside	0.54	0.66	0.59
Hit Object	0.75	0.07	0.12
Overturned	NA	0	NA
Vehicle/Pedestrian	0.88	0	0
Other	NA	0	NA

Difficulties

- Data skewed towards one class
  - Results in always predicting one category
  - Low overall precision, recall
  - Better model might implement boosting
- Implementing C4.5 from ground up
  - Calculating entropy and information gain
  - Using RDD's instead of DataFrames

Thank You!